AZ EMBERI VISELKEDÉS DINAMIKÁJÁNAK SZÁMÍTÓGÉPES VIZSGÁLATA

COMPUTATIONAL STUDIES OF HUMAN DYNAMICS

Az MTA Doktori cím elnyeréséhez készített disszertáció

DR. KARSAI MÁRTON

Rényi Alfréd Matematikai Kutatóintézet

Hálózat és Adattudományi Tanszék
Central European University

2022

First printing, August 2022

Contents

Acknowledgement vii

List of symbols ix

1 Introduction, positioning and terminology 1

1.1 Introduction **1**

1.1.1 Scientific landscape 1

1.1.2 Methodological challenges 3

1.1.3 Positioning and outline 7

1.2 General concepts and terminology **8**

1.2.1 Complex networks 8

1.2.2 Social networks 10

1.3 Datasets **11**

1.3.1 Data representation 11

1.3.2 Datasets in hand 12

2 Bursty Human Dynamics 15

2.1 Introduction **15**

2.2 Characterisation of bursty phenomena **17**

2.2.1 The Bursty Train Size Distribution 20

2.3 Cyclic patterns in human dynamics **23**

2.4 Observation of bursty phenomena **24**

2.4.1 Bursty egocentric network evolution 25

2.4.2 Bursty communication in egocentric networks 27

2.5 Models of bursty human phenomena **29**
2.5.1 Model of individual bursty dynamics with event trains 29
2.5.2 Model of communication balance of dyadic event trains 32
2.6 Conclusions **34**

3 Temporal Networks 35
3.1 Introduction **35**
3.1.1 Static vs. Temporal Networks 35
3.1.2 Time-scales of network dynamics 36
3.1.3 Representations of temporal networks 37
3.1.4 Some characteristic measures of temporal networks 38
3.2 System level characterisation **39**
3.2.1 Aggregation time for social communication networks 39
3.2.2 Entropy of Dynamical Social Networks 41
3.3 Random Reference Models **43**
3.3.1 Naming and and organisation of RRMs 44
3.3.2 Modelling concept and demonstration of RRMs 44
3.4 Higher order representations **48**
3.4.1 Temporal motifs 48
3.4.2 Weighted event graphs 53
3.5 Generative models of temporal networks **59**
3.5.1 Activity-driven network model 59
3.5.2 Memory processes in egocentric network formation 60
3.5.3 Individual heterogeneities in social capacity 63
3.5.4 Dyadic closure and node removal mechanisms 65
3.6 Conclusions **68**

4 Collective phenomena on networks 69
4.1 Introduction **69**
4.2 Static observations of collective social phenomena **70**
4.3 Dynamic observations of social spreading phenomena **71**
4.3.1 Complex contagion process in spreading of online innovation 71
4.3.2 Local cascades induced global contagion 74
4.3.3 Social experiments and behavioral data collection 77
4.4 Modelling simple spreading phenomena **79**
4.4.1 Simple contagion processes 80
4.4.2 Controlling contagion processes in time-varying networks 81
4.4.3 Link transmission centrality in large-scale social networks 84
4.4.4 Switchover phenomenon induced by epidemic seeding on geometric networks 87
4.5 Modelling complex spreading phenomena **92**
4.5.1 Complex contagion processes 93
4.5.2 Dynamical threshold model with immune nodes 94
4.5.3 Threshold driven contagion on weighted networks 99
4.5.4 Threshold driven contagion on multiplex networks 102
4.5.5 Threshold driven contagion on temporal networks 105
4.6 Conclusion **106**

5 Discussion 107
5.1 Perspectives 107
5.1.1 Future directions of my research 107
5.1.2 Perspectives of my associated fields 109
5.2 Conclusions 109

Bibliography 111

Acknowledgement

Over the years I had the opportunity to work with several exceptional people who truly shaped my opinion, sophisticated my scientific approach, and set high standards to follow in scientific and personal manner. First of all I would like to acknowledge my mentors who helped me to develop, and to reach opportunities since the beginning of my scientific career. Pr. Ferenc Iglói and Dr. Jean-Christian Anglès d'Auriac were my PhD supervisors who set me on the path towards my scientific journey. Pr. Kimmo Kaski, Pr. Jari Saramäki and Pr. János Kertész were my mentors at Aalto University who introduced me to new scientific challenges, which are still in the core of my research. They opened me the door for new opportunities, which entirely changed my professional motivations and goals. Their continuous encouragement, patient, stimulating suggestions and healthy criticism helped me greatly to develop my scientific identity. I am especially beholden to Pr. Alessandro Vespignani at Northeastern University, who first of all provided me the unconditioned opportunity to explore my own scientific path, and who set me an example of such high professional and personal standards, that I can use as a compass to follow during the rest of my career. I am truly grateful to Pr. Eric Fleury, who welcomed and integrated me at ENS Lyon. His continuous support helped me to grow in every sense. He trained me on how to propose and manage projects, how to supervise students and a team, while not loosing connection with the core of scientific problems and with the fun they provide in life. Finally, I am truly thankful for Pr. János Kertész for his support not only as a mentor but to direct me towards the possibility to join the Department of Network and Data Science at the Central European University, where I found my scientific home.

I regard science as a social effort and thus I have built several collaborations with excellent researchers. Exploring problems together with them helped me to develop ideas without compromises, while in turn many of them became my friends. They are J. I. Alvarez-Hamelin, A.-L. Barabási, A. Barrat, F. Battiston, S. Bernhardsson, G. Bianconi, P. Bíró, V. Blondel, E. Bokányi, H. Bouchet, R. Burioni, G. Cencetti, J-P. Chevrot, J-P. Cointet, C. Cherifi, H. Cherifi, V. Colizza, D. Czifra, W. Du, A. Flamini, L. Gauvin, M. Génois, J.P. Gleeson, P. Jensen, H.-H. Jo, S. Juhász, J. Karikoski, R. Kikas, M. Kivelä, J. Koltai, J. Komjáthy, CS. Kolok, L. Kovanen, G. Krings, K. Kutasi, B. Lengyel, B. Lepri, S. Liu, L. Lovász, J-P. Magué, F. Menzer, P. Mercklé, M. Minnoni, D. Mocanu, A.B. Modiri, M. Musolesi, A. Nardy, H. Torres Marques Neto, R. K. Pan, N. Perra,

W. Quattrociocchi, A. K. Rizi, G. Röst, L. Rossi, C. Roth, Z. Ruan, B. Ságvári, C. Sarraute, F. Schweitzer, K. Sun, Á. Szabó-Morvai, T. Takaguchi, M. Toricelli, C. J. Tessone , G. Tibely, M. Tizzoni, M. V. Tomasello, E. Ubaldi, E. Valdano, O. Vásárhelyi, C. L. Vestergaard, A. Vezzani, L. Weng, Q. Zhang, and K. Zhao.

Transition from being a student to supervise students is a journey I try to make every day. That is why I am very grateful for all students and postdocs I have had the chance to work with, for their open minded approach and their susceptibility and commitment to explore new ideas. They are L. Alessandretti, G. Laurent, J. Cambe, A. Baltzer, J. Faria de Oliveira, M. Morini, Y. Leo, S. Unicomb, J. Levy Abitbol, S. Dai, R. Hilman, L. Napoli, E. Andres, A. Manna, H. Hours, Y. Liao, S. Lerique, L. Espín, I. Iacopini, G. Ódor, and R. Vaudaine.

I also would like to acknowledge the support and friendly environment created by the DANTE research team, the IXXI and LIP laboratories, and the Computer Science Department at ENS Lyon. Namely, I would like to mention P. Gonçalves, P. Nain, A. Busson, T. Begin, C. Crespelle, I. Guérin Lassous, R. Gribonval, L. Lecot, N. Trotignon, P. Borgnat, S. Thomasse, and D. Stehlé. Most of all, I am grateful for my colleagues E. Omodei, P. Kralj Novak, M. Pósfai, F. Battiston, G. Iñiguez, T. Peixoto, B. Vedres, O. Peredi, and J. Kertész at DNDS CEU who provide me with an inspiring, professional and friendly research environment where I can work and teach without compromises. I am also very thankful for my colleagues at the Rényi Institute, L. Lovász, D. Miklós, A. Stipsicz, M. Szomolányi and E. Szakonyi who supported my research in every possible ways over the most recent years.

I am indebted to my parents and overly thankful for my brother, my sister and their families, who support and encourage me until today despite the large geographic distance between us.

Finally and most importantly, I would like to express my deepest gratitude to my little family. My wife Sophie and my children Lilla and Ábel are the most important things in my life. Their unconditional love and trust provide me an invaluable support; any of these achievement would have been simply impossible without them.

List of symbols

This is a non-complete list of acronyms, which only collects the most frequently used notations in the text. Some notation specific to a given study will be introduced at the place of discussion, this way allowing the multiple usage of the same symbols.

G	graph
V	set of nodes of a graph
E	set of links of a graph; bursty train size
$e_{u,v}$	link between nodes u and v
N	size, i.e., number of nodes in a network
M	number of links in a network
$w_{u,v}$; w	weight of the link $e_{u,v}$; link weight in general
G_v	egocentric network of node v
V_v	neighbour set of node v
E_v	set of links in G_v
G_t	temporal network
E_t	set of events (temporal events)
$G_{[0,t]}$	aggregated temporal network
D	event graph
E_D	set of links in D
k_v; k	degree of node v; node degree in general
k_{in}; k_{out}	in and out-degree of a node
$\langle k \rangle$; z	average degree
s_v; s	strength of node v; node strength in general
C_v	clustering coefficient of a node v
C	average local (or global) clustering coefficient

$O_{u,v}$; O	overlap of link $e_{u,v}$; link overlap in general
$n_{u,v}$	number of common neighbours of nodes u and v
ρ	network density
$P(x)$	Probability density function of a variable x
$\langle x \rangle$	average value of variable x
γ	degree distribution exponent
$d(u,v)$	graph distance between nodes u and v
$BC(u)$	betweenness centrality of node u
T	observation time period
$A_i^{ev,ac}$	An attribute i of an event ev or action ac
A_j^v	A meta-data attribute of an individual v.
t; t_i	time; time of the ith event.
$x(t)$; $ev(t_i)$	binary event sequence; sequence of event timings
τ	inter-event time
τ_r	residual time
α	inter-event time distribution exponent
$A(t_d)$; t_d	autocorrelation function, delay time
γ	autocorrelation exponent
β	train size distribution exponent
ω	event frequency
B	bursty parameter
$p(n)$	memory function
$\mathbf{k}$, $\mathbf{w}$	sequence of degrees, weights, etc.
$\langle \ell \rangle$	average path length
Δt	maximum time between causal events
a_i	activity of node i
φ	individual adoption threshold
Φ	integer adoption threshold
ϕ	average adoption threshold
r	fraction of blocked nodes for adoption

1. Introduction, positioning and terminology

1.1 Introduction

The goal of this thesis is to summarise my scientific achievements since my PhD in 2009 and to position these contributions on the scientific landscape. My scientific motivations have always been *curiosity driven* and focusing on answers to the question "*why?*" (rather than to develop novel methods or technologies only). This attitude may be rooted in my training in physics but certainly resulted in a very heterogeneous scientific portfolio with projects and publications ranging between and combining knowledge from several domains. However, these results all have a common ground to address some aspects of human dynamics using computational methods and concepts rooted in physics. It builds on the conventional results and recent developments of several fields and as such it is truly interdisciplinary, which makes it difficult to fit into the traditional categories of academic disciplines. However, I judge this scientific multi-pluralism as an advantage, as it provides a ground for the recombination of knowledge of foreign disciplines to solve unconventional problems.

1.1.1 Scientific landscape

My thesis is naturally related to *physics of complex system*, which studies the emergence of collective phenomena that arise from the interactions of entities in many-agent multivariate systems. The emergence of collective behaviour has been historically studied in the framework of statistical physics [329] of phase transitions and critical phenomena. My PhD landed within these fields as it is mostly about phase-transitions of cooperative behaviour emerging in modelled physical systems like in spin models, diffusion, or percolation phenomena [201, 193, 196, 197, 194]. Meanwhile, collective phenomena occur not only within the realm of the physics of matter, but also in many other areas, including biological, social, and economic systems. The global spreading of a pandemic, the emergence of social movements, the collective migration of animals, the growth of tumours, or the coordinated firing of neurons are all examples of interdependence and collective behaviour [327]. However, advances in these areas have been hampered by difficulties in collecting the vast amounts of detailed data necessary for validating theories and developing quantitative approaches. Especially in the social systems arena, the community has confronted the insurmountable obstacle of a lack of data on human behaviour at multiple scales. Consequently, in many ways, it is currently easier to

observe tiny bacteria or galaxies light years away than our fellow humans [363].

These foundational limitations are now being obliterated by the *digital data revolution* [260, 79]. Remarkably, every 1.2 years, more human-driven socioeconomic data are produced than during all preceding years of human history combined. Finally, we are in the position to follow the evolution of large real-world systems and detect the emergence of collective social behaviour "in vivo". This boom in data collection was induced by widely adopted new technologies, such as mobile phones or online services, and was escorted by the development of new computational designs. Recent availability of *high performance computational resources*, the advances in *social computation methods*, and the advent of *advanced machine learning and statistical data analysis* all contributed to the success of data-driven research of human behaviour. However, at the same time, such studies highlight our limitations in knowledge about complexity itself and challenge us on the foundational aspects (such as conceptual, theoretical, and modelling issues) that form the basis of our understanding of complex systems.

In this context, *network science* has been assigned an increasingly relevant role in defining a conceptual framework for the analysis of complex systems (as discussed in details in Section 1.2.1). Network science is rooted in statistical physics and graph theory and it is concerned with structures that map real entities and their interactions to graph nodes and links [370, 40, 18, 280] to describe the architecture of complex systems. For a long time, this mathematical abstraction has contributed to the understanding of real-world systems in physics, biology, chemistry, social sciences, and economics. Recently, however, the enormous amounts of detailed data, electronically collected and meticulously catalogued, which finally become available for scientific analysis and study, led us to the discovery that most networks describing real world systems show the presence of complex properties and heterogeneities, which cannot be neglected in their topological and dynamical description. This has called for a major effort in developing the methodology to characterise complex networks [278, 173, 108, 388, 226], to describe the observed structural and temporal heterogeneities [18, 173, 39, 200], to capture the multilayer nature of connectedness [216], to detect and measure emerging community structure [126], to identify the effects of spatial embeddedness [46], and higher-order structural [384, 150, 49] and temporal [195, 218, 268, 223, 217] correlations determining the emerging network structure, etc. All these efforts have brought us to a point where the science of complex networks has become advanced enough to help us to disclose the deeper roles of complexity and gain understanding about the behaviour of very complicated systems like global epidemic, transportation systems, the brain, or society.

Finally, we are able to tackle challenges, which were not addressable earlier due to lack of data, but now they are possible through the quantitative observations of the social behaviour of individuals and groups in global settings. These advancements called for the emergence of the new field of *computational social science* (CSS) [234] with the aim to develop the methodology for the quantitative description and modelling of social systems. Beyond topical relevance, any question of CSS translates to computational problems, and many of them can be translated to concepts developed in physics of complex systems. Moreover, CSS is based on methodologies borrowed from conventional social sciences, cognitive and behavioural sciences, psychology, statistics, physics, computer science, and network science. Direct applications of methodologies from this broad set of disciplines were proofed to be successful at the outset, but at the same time they inherited concepts, which potentially mislead the description of the systems in focus. This way, for the quantitative description of human behaviour, the development of entirely new concepts were called for, leading to the emergence of domains like computational economics, human dynamics, social simulations, or computational linguistic, etc. Potentials of these advancements are not only to enhance data-driven reasoning in explaining social phenomena [363] but to fuel a paradigm shift to introduce social sciences as more quantitative fields.

Dynamical processes like information or infection spreading, the adoption dynamics of inno-

vations, memes, fads, or opinions can be effectively studied through the adoption of the network picture and methods from the physics of collective behaviour. Central question here is how the structure of interactions effects the critical behaviour and phase transitions, commonly characterising such phenomena [43]. The theoretical description and synthetic modelling of these systems have been earlier studied with tools borrowed from statistical physics, critical phenomena and computational modelling, while *data-driven modelling* [317, 195] and *statistical learning* [127] are recent promising directions to bring predictive modelling of dynamical processes closer to real observations. It has been shown that several systems are crucially influenced by the heterogenous number and strength of interactions as well as by the time-varying nature, multiplexity, and structural and temporal correlations of them. An approach from physics is naturally helpful to address these problematics by quantitatively studying any collective social phenomena and in a broader sense dynamical processes, which evolve on social networks and are driven by information (e.g. influence, rumours, or memes) transmitted via social interactions.

Human dynamics [325], on one hand, concentrates on individuals in terms of their actions, autonomous behaviour, mobility and migration, opinion formation, or spatio-temporal dynamics in geographic or mental spaces, etc. However, even at this level of description, the ego[1] in focus is always embedded in a socio-economic environment, which cannot be ignored. On the other hand, human dynamics is concerned also with dyadic and egocentric interaction dynamics to unveil temporal patterns of interactions, mesoscopic group formation, inter and intra-community dynamics in social networks. On the system level, it addresses the dynamical emergence of the social networks and any global human dynamical phenomena, like patterns of collective motion, emergence of collective movements, spreading of opinion, products or information, etc. This way, it can be identified as a sub-field of computational social science even its general focus is somewhat broader than to address sole social behaviour. It is also strongly entangled with network science, physics of dynamical systems, data science, computational science and other fields mentioned above.

Human dynamics as a field originates from direct observations of individual human behaviour [39, 148]. In this thesis, I propose a broader interpretation of the field, by considering not only to directly study dynamical human behaviour, but also its consequences [200]. This way, I include problems under this umbrella, which are conventionally associated to other fields, but their understanding is crucially depending on some aspects of human dynamics.

1.1.2 Methodological challenges

The methodological steps in research of human dynamics basically follow the conventional epistemological structure common in physics and in natural sciences, but with innovative, topic relevant techniques applied at each stages. Next, I give a short and certainly non-complete summary about the methods, techniques and paradigms applied at each phases, to obtain generalisable and verifiable knowledge from data-driven observations of human behaviour.

Data collection

Digital behavioural data has the advantage to record human behaviour in its own settings, in fine details, potentially accompanied with spatial, temporal and demographic details, for millions of individuals, without the common observational biases [79]. This way, it may provide statistically more significant, variant, and better generalisable observations about human behaviour. This is opposed to census data where individual details are aggregated, or to earlier designs of experimental social studies where observations were typically made in artificial settings involving a handful of participants. Digital behavioural and especially relational (network) data can be collected from

[1] Note that naming a person an individual or an ego are deemed equivalent in this thesis, as they may be conventional in different terminologies, but they refer to the same, a person.

various sources. *Life-logs* and *individual tracking* [113, 181, 334] relies on recent technologies like mobile phones or radio-frequency IDs (RFID) [44, 252] to follow people with their consent. Such projects provide temporally detailed records about a limited number of participants potentially including their whereabouts, communication (email, mobile call, short messages, online social networks,...), physical proximity, application usage, sleeping times, circadian patterns, just to mention a few examples. Other great data sources are provided by the automatically collected communication records for archiving or billing purposes, like *emails* [116] or records of *mobile phone communications* [54]. Such data usually record temporally detailed lists of dyadic communication actions of millions of anonymised users (e.g. customers of the same mobile provider), potentially with their actual location and length of interaction, but certainly without any information about the content of their communication. Typically they are accompanied with meta-data sets including demographic and socio-economic details about the users. The recent popularity of *online social services* contributed enormously to this digital data endeavour. Services like Facebook, Google, Twitter, Spotify or gaming companies all record and store detailed personal and service usage informations about their customers and use them for development, marketing and advertisement purposes. At the same time, they share data via open APIs or data challenges, which in turn are extremely useful for external developers and scientists [373, 282]. Finally, *online collaborative platforms* [187] and *archiving services* [2] record relational data about professional interactions like scientific collaborations or project developments. In addition, there are myriad other ways to collect and access digital data valuable for research, and this list will ever expand due to the emergence of new technologies and online services.

Some major disadvantages are also rooted in the data-driven approach to human behaviour. First of all, such digital datasets are more like field experiments, they are not collected in a controlled experimental setting and this way they are *not reproducible* at the finest level of details. They are somewhat similar to astronomic observations, where despite recurrent patterns and regularities, the observation of the Universe is unique at any point in time. Possible solutions to these shortcomings are provided by the recent and more-and-more frequently used designs of online social experiments [309], which are controllable, scalable, and relatively un-expensive. Depending on the problem setting, they may employ unique experimental designs [76] or use online survey platforms like Amazon Mechanical Turk [1, 69] or SurveyMonkey [9]. There are also some *demographic biases* commonly characterising digital behavioural datasets. The digital data collection about individuals is conditional to the usage of digital devices (like mobile phones or computers) and communication services (email, online social networks, etc.). This way some demographic groups like elderly people, ones living at rural areas, or people with smaller income may appear under-represented in the data. However, due to the development of the telecommunication infrastructure, the increasing availability of smart devices, and the ever seen popularity of online services, these biases are vanishing even in developing countries, yet they are necessary to be noted. Another disadvantage of the actual digital data practice has been raised on the ethical side. The automatic collection of *highly sensitive personal data* and the potential tracking and identification of people without their consent have been identified as major concerns. Customers using free online services, unintentionally pay with their data, which in turn is used and re-sold for marketing and other purposes. Recent developments in data ethic and personal information treatment has been enforced by policy makers to put the rights to own data to the hand of the customers, this way closing the ethical gap between practice and privacy [4]. However, this also indicates challenges in research and development, to design appropriate standards of data privacy and methods for data collection, which aggregate individual details on a necessary level to avoid potential re-identification [102], but yet to contain valuable information for scientific investigations.

Observation and analysis

Hundreds years old methodologies provided by *statistics* were developed for the analysis of conventional social and economic data, where the usual obstacle was the small sample size. Due to this limitation advanced statistical models, measures of variance and confidence were developed to be able to state something meaningful from small data and to identify biases [333]. These techniques were inherited to the contemporary methods of *statistical data analysis* [290], where the sample size is usually not the bottleneck anymore, but the heterogeneity of the observed quantities raises new challenges. To extract information from large digital datasets, the new field of *data mining* [379] grew out on the basis of *statistics*, *machine learning*, and *data bases* to identify patterns from large data and transform them into understandable knowledge. The methodology of this field circles around data retrieval, pre-processing, transformation, data mining, and interpretation and aims to solve problems like anomaly detection, association learning, clustering and classification, regression and inference.

Network analysis provides other tools, which were found useful to learn about the dynamics of humans, embedded in a social, economical, geographical or abstract spaces [367]. Building on tools originated from *graph theory* and *statistical physics*, network analysis operates with measures characterising nodes and links, local and mesoscopic structures and the entire connected network to identify locally and globally important vertices and relationships, assortative correlations, communities, core-peripheral structures, and heterogeneous statistical characteristic of the structure, and many others [278]. On the other hand, network analysis provides several ways to represent relational data as network structures with increasing level of complexity. Beyond the simple static, weighted and directed description of interactions, recently it grounded the methodology to take into account the evolution of networks, the time-varying nature of interactions [173], the multiplexity of links associated with several types [216, 55], the spatial embeddedness of nodes [46], or several attributes assigned to nodes and links. In addition, *randomised reference models* [195, 134] provide further analysis tools to identify significant structural and temporal patterns of interactions via the controlled shuffling of the temporal and structural informations encoded in the network.

Modelling

Modelling systems of human dynamics borrowing concepts from statistical physics, complex systems, machine learning, and computational social science, with the aim of better understanding and predicting human behaviour. One can identify three main branches in this endeavour:

Statistical models are common in main stream empirical social science to detect statistically significant correlations between a variable describing a social phenomenon and a variable thought to explain it [172, 367]. This approach is originated in advanced statistical methods, however recently, it applies technics borrowed from machine learning, data mining, and Bayesian learning. The main advantages of this approach are to account for individual differences to identify groups of egos with similar behaviour and to make statistical predictions even on the single ego level. Moreover, these methods allow the detection of causal inferences in data [130], while crucially they do not address mechanistic processes driving human behaviour. In this way they may help to build hypothesis of social behaviour but they have limited capacities in proof-of-concept and scenario-testing type of modelling.

Mechanistic models are based on concepts borrowed from physics, where egos are identified as interacting entities with decisions driven by presumed mechanisms [172, 163]. They typically model emergent behaviour whereby larger complex phenomena arise through interactions among smaller or simpler entities such that the emergent system exhibits properties the smaller entities do not. These generative processes are sometimes formalised into coupled dynamical differential equations with analytical or numerical solutions, and emulated in the realm of social simulations [136] as large-scale computer simulations where the stochastic nature of human behaviour is considered by probabilities of possible decisions at the individual level. School examples are generative

network models [61, 206] employing the minimal but sufficient set of microscopic mechanisms, e.g. preferential attachment, memory, or reinforcement, which lead to some emerging property of the global network structure, like heterogeneous degrees or community structure. These models can consider social mechanisms influencing the decisions of egos, they intrinsically take into account causal relationships in reasoning, and provide ways to understand emerging phenomena in social systems [73]. On the other hand, they usually assume that egos are identical and statistically equivalent entities (above the cognitive level) and can be characterised by their "average" behaviour. This approach has been proofed to be successful to simulate system level behaviour, and it provides an ideal test bed for hypothesis and scenario-testing. On the other hand, as it usually neglects individual heterogeneities, thus it has limitation for a fine-grained description of a social system.

Data-driven models is a new branch in modelling social systems and co-evolving dynamical processes [340, 195]. They integrate real-world data with synthetic models to concentrate on the effect of selected mechanisms, while keeping the rest of the system as close to reality as possible. Data can enter a synthetic model in several ways, e.g. by parameters determined from empirical systems [340], by replacing some parts or the entire system with real data and only model the effects of selected mechanisms [198], or even by replacing real entities in an observed population by model agents to observe their simulated behaviour in a real setting [132, 72]. In addition, to better understand simulated dynamical processes evolving on real networks , the effects of structural and temporal correlations can be testified by using *randomised reference models* [134, 195, 218].

Although these three modelling paradigms have relevance on their own, they provide complementary advantages for better understanding. To understand, model, and predict social systems, after observation it is crucial to ask the question "why?". Using mechanistic and data-driven models one can build and demonstrate hypothesis(es) by identifying the minimal set of causal mechanisms leading to the emergent behaviour in focus. This knowledge may provide relevant information to select features to train statistical models for prediction. Applying models alone from these three paradigms would leave us (a) with self-serving mechanistic models with multiple hypothesises explaining the same phenomena, and (b) with simple rules-of-thumb like black-box experiments using statistical models.

Verification and validation

Verification and validation of models of human dynamics are difficult as social systems are far from being deterministic and as such they do not provide any 'ground-truth' (unlike systems in physics) for formal evaluation of modelling results [158]. Nevertheless, various techniques exist to verify results provided by models of one of the three approaches we discussed above. In case of statistical learning, datasets are commonly divided for training and validation sets. After training a model on the former one, verification standards are defined via the *sensitivity* (true positive) and *specificity* (true negative) recalls from an independent validation set [290, 379]. These measures are able to quantify the proportion of correctly recovered positive and negative attempts of the predicted features even on the individual level and can help to identify deterministic independent features for prediction. However, since they are simply considering the predicted output of a black-box experiment they are unable to proof dependencies beyond the phenomenological level.

Verification of generative models is impossible on the level of single entities but on the level of the emergent phenomena. This can be done by directly comparing the solution of the model's formal description to its simulated dynamics or stationary state [43], or by verifying the emergence of some expected characters on the system level [278]. Data-driven models are typically compared to real world observations, whether directly or by feeding the modelled process with real parameters tp compare predictions to empirical outcomes [340]. Moreover, biases induced by complex mechanisms can be verified by data-driven simulations by comparing their outcome to corresponding simplified synthetic processes.

Applications

Beyond scientific merit and understanding, human dynamics has far-reaching applications in several domains. First of all results of statistical modelling provides functional knowledge for predictions and inference of e.g. individual communications, mobility, or interactions patterns or online and offline behavioural dynamics. More concrete applications and algorithmic solutions of data mining results on human dynamics are usually disseminated on KDD conferences [8]. In terms of generative and data-driven modelling applications are also manifold. The prediction and scenario testing of crowd dynamics, cooperative behaviour, traffic congestion, public transportation, adoption of products and services, or the forecast of global pandemic and diffusion of information, memes, trends and opinion are only a few examples, which rely on data, observations, and modelling of human dynamical systems.

This way, human dynamics as a field has an un-questionable and increasingly relevant role in understanding human behaviour embedded in modern techno-social societies [363]. It contributes to the application of the concepts of physics, outside of its conventional fields, and helps the transformation of social sciences to become a more quantitative field. It pushes the frontier of complex systems science in terms of better understanding collective phenomena, and computer science through the advancement of data science and computation, and it fuels the paradigm shift introduced by digital data in science, technology and several application domains.

1.1.3 Positioning and outline

As I explained in the introduction, my research is very heterogeneous and builds on several disciplines. If I would need to summarise in short, I am a physicist trained network scientist who is using methods from statistical physics, complex systems, and computer science to answer questions about emergent phenomena in general, and more generally in human dynamics. My contributions include theoretical works, methodological contributions, data collection, analysis and modelling studies to push forward the fields of network science, physics of complex systems, data science and computational social science. My research lands in the field of *network science*, which stems from the roots of statistical physics and graph theory. On the fundamental side of physics, my contributions are related to the physics of complex systems at large, to study network dynamics and collective phenomena with the concepts of phase transitions and critical phenomena, dynamical processes, percolation theory, computational physics and Monte Carlo methods and agent-based modelling. On the applied side of physics my results contributes to modelling of biological and social spreading processes, dynamical systems in terms of the structure and dynamics of human interaction, socioeconomic, and mobility networks and ongoing dynamical processes.

Outline of the thesis

Although my works are at the intersection of various domains they all address emergent behaviour in systems of human dynamics. Most of them can contribute to three main topics, which also defines the overall structure of this thesis.

After this short introduction, in the remaining of the first Chapter I am going to introduce the general definitions and concepts in complex networks, together with some theories and general properties of social networks, which will be used throughout the whole thesis. After that I give short summaries of various datasets what I used during the last years, and which are recurrent in several studies discussed later. Following that there are three scientific chapters, each starting with a short literature review followed by the summary of my contributions to these domains. In Chapter 2, I discuss bursty dynamical systems, which is the first main topic of my contributions. This Chapter builds on a set of studies, in which I proposed novel methodologies to characterise and model heterogeneous temporal behaviour of individuals and once they are connected in a network structure. The second main topic of my research is discussed in Chapter 3. There I summarise

my theoretical and modelling contributions to the foundation of temporal networks. These works circle around new theoretical methods to interpret and characterise temporal networks at the micro-, meso-, and macroscopic scales. In addition, I propose the observations of various mechanisms and two modelling frameworks to explain the emergent properties of temporal structures. In Chapter 4 I address my works on dynamical processes evolving on static or temporal networks. These studies investigate the structural and temporal correlations, which control emergent collective phenomena like spreading processes. I will address simple contagion models on real and modelled temporal networks, and complex contagion processes, especially the empirical observation and modelling of social spreading phenomena. Finally, in Chapter 5 I will summarise my contributions and will draw my motivations for the future.

1.2 General concepts and terminology

In this Section, I briefly introduce some general definitions and concepts, which will be used throughout the whole thesis. All the definitions are related to complex networks, first introducing general characteristics, while second some specific concepts, which concerns social networks. This list of definitions, however, is far from being complete to give a thorough introduction to the field, while some specific definitions will be introduced at the related Sections in the following Chapters.

1.2.1 Complex networks

Networks provide a way to map the architecture of complex systems by identifying interactions between their interconnected components. A network is commonly interpreted as a graph defined by a set $G = (V, E)$. Here V defines the set of $v \in V$ nodes representing components, and $E \subseteq V \times V$ is the set of $e_{u,v} = (u, v) \in E$ links, which indicate if two components $u, v \in V$ are interacting in the system[2]. The cardinality of these two sets assign the size $N = |V|$ and the number of links $M = |E|$ of the network. In their first approximation, networks are defined as simple graphs, thus they are undirected (do not distinguish between links (u, v) and (v, u)), while assume no self-loops (v, v), or multiple connections between the same pairs of nodes. On the other hand several extension of this representation is possible. For example, if we assign an orientation to each link, i.e., we distinguish between (u, v) and (v, u), we arrive to the definition of *directed networks*, while *weighted networks*, defined as a triplet $G = (V, E, w)$, assume a weight function $w : (u, v) \to \mathbb{R}$, associating a number (or another quality) to each link to capture, e.g., the interaction strength between connected entities.

Networks can be characterised on various topological scales. To describe them on the local scale, i.e., at the level of nodes and links, we consider the $G_v = (V_v, E_v)$ *egocentric network* of each node v, which includes the central node, its first neighbours, and all links among them. At this level the most important character of a network is the *node degree*. The degree k_v of a node $v \in V$ is the number of links it is incident to, or in other words its number of neighbours[3]. Formally it can be defined as $k_v = \sum_{u \in V \setminus \{v\}} \delta(u, v)$, where $k_v \in [0, N-1]$ and $\delta(u, v) = 1$ if $(u, v) \in E$ and 0 otherwise. Similarly, we can define the k_{in} in-degree and k_{out} out-degree of a node in directed networks, counting the number of its incoming and outgoing links. In case of weighted networks, the equivalent quantity is called the weighted degree or *node strength*, defined as $s_v = \sum_{u \in V \setminus \{v\}} w(u, v)$, i.e. the sum of weights of adjacent links to a node v. Another important node characteristic is the *local clustering coefficient*. It quantifies the local connectedness via counting the fraction of closed triangles (i.e., three nodes connected by three links) in the egocentric network of a node v. Formally it is defined as $C_v = \frac{2|\{e_{v,w} : v,w \in V_v, e_{v,w} \in E\}|}{k_v(k_v-1)}$, where the denominator $(k_v(k_v-1))/2$ counts the maximum

[2] Note that several naming conventions of the same objects are accepted in different disciplines. Networks are also called structures or simply graphs, nodes are equivalently named as sites, components or vertices, while links can be called as edges or ties. In this thesis I treat equivalent the corresponding names.

[3] Note that although k_v is a specific character of any node in V, later we may use the simpler notation k, which assigns degree in general of a node in the network. Similar convention will be used for any other node property.

number of connections between the neighbours of a node with degree k_v. The local clustering coefficient is conventionally used in social network research [370] where its network average value, $C = \sum_{v \in V} C_v / N$, characterises the average local connectedness of the structure. A similar quantity can be defined for links, called the *link overlap* [155, 288], which measures the fraction of common neighbours of connected nodes. It is formally introduced as $O_{u,v} = \frac{n_{u,v}}{(k_u - 1) + (k_v - 1) - n_{u,v}}$, where $n_{u,v} = |\{V_v \cap V_u\}|$ is the number of common neighbours of nodes u and v.

On the global scale, networks are commonly characterised by the distributions and statistical means of different local quantities. In the following we define some of them for simple networks (undirected with no loops and multiple links), but all measures can be generalised for more complicated representations. The simplest global measure is *network density*, $\rho = \frac{2|E|}{N(N-1)}$, which is the fraction of present links as compared to the possible number of links in the network. A more informative measure is the probability density function of node degrees, also called the *degree distribution*, defined as $P(k) = \frac{1}{N}|\{v \in V, k_v = k\}|$, and its mean $\langle k \rangle = \frac{1}{N}\sum_{v \in V} k_v$ called the *average degree*. This distribution determines the probability that a randomly selected node in the network has degree k, while its mean gives the average number of neighbours of a node. In real-world networks, $P(k)$ usually appears as a broad distribution with a long tail ranging over several orders of magnitude [18], indicating present high degree heterogeneities. Such distribution are commonly approximated with a power-law distribution scaling as $P(k) \sim k^{-\gamma}$, with a characteristic exponent γ. Although this approximation has been proven to be useful, the $P(k)$ of several real-world networks may be better fitted by other broad statistical distributions, like lognormal or stretched exponential, as it is summarised in [82]. Note that several other characters of real world networks are distributed heterogeneously, like link weights or node strengths, and can be similarly characterised. *Global clustering coefficient* is another global scale measure, which quantifies the fraction of closed triangles on the network level. It is defined as $C = \frac{3 \times |Triangles|}{|Triplets|}$, where triplets are subgraphs of three nodes connected by two links[4]. Note that the weighted average of the local clustering coefficient is equivalent to the value of the global clustering coefficient.

Paths in networks are very important as they determine the possible routes of information flow between nodes, and code the global connectedness of the structure. Two nodes, u and v, in a network are connected by a path if there exist a sequence of links, $(u,a),(a,b),\ldots,(x,y),(y,v)$, which starts from u, ends at v, and each consecutive links share one common node. The *length of a path* is the number of links in the corresponding sequence, while the *distance* $d(u,v)$ between two nodes is given by the length of the shortest path, i.e, the shortest sequence of links, which form a path between them. Shortest paths can be used to define sensitive measures of centrality and overall connectedness of a network. First of all, the *average distance*, $\langle d \rangle = \frac{1}{N(N-1)}\sum_{u,v \in V} d(u,v)$, captures the average number of steps required to pass information between any pairs of nodes. Another measure is the *network diameter*, defined as $\max_{u,v \in V} d(u,v)$, gives the maximum of any distances between any pairs of nodes. *Betweenness centrality* is a path based measure to identify central nodes in the network, which lays on several shortest paths thus keeping the structure connected, or controlling effectively information diffusing on the network. It is defined as $BC(v) = \sum_{v \neq s \neq v in V} \frac{\sigma_{u,v}(s)}{\sigma_{u,v}}$, where $\sigma_{u,v}$ is the number of shortest paths between nodes u and v, while $\sigma_{u,v}(s)$ is the number of those which run through node s. Existing paths also determine the connected components of a network. A connected component contains a set of nodes, which are all available from each other via paths in the network. The largest of them, called the *giant connected component* (GCC) or the *largest connected component* (LCC), plays a special role as its size characterise the overall connectedness of the network and determines the maximum size of any macroscopic phenomena emergent in the system. The size of the *second largest component* (or the

[4] I do not distinguish in notation between the average local clustering coefficient and the global clustering coefficient. As default, in this thesis C will refer to the former one if not noted otherwise.

average size of components other than LCC) is also important in some cases. By varying the number of links in the network, its connectedness can be interpreted as a *percolation process* [16] with a phase transition point, where the size of the LCC (commonly identified as the order parameter) vanishes and the size of the second LCC (identified as a susceptibility like character) shows a singularity in the thermodynamical limit.

1.2.2 Social networks

As most of the works summarised in this thesis concentrate on social systems, it is useful to introduce a non exhaustive list of some commonly accepted concepts and general characters of social networks to help the reader. In social networks nodes are associated to individuals, while edges represent social ties between them. Social networks are inherently *dynamical* on multiple temporal scales, as people enter and leave the system and create and break social relationships, and maintain them via rapid communication events. Social ties can be manifold by distinguishing between family, friend, intimate, professional or service types of relationships. Such variety of interactions can be represented via *multilayer* structures with layers associated to social ties of different types. Social networks are also *embedded* in geographical space due to the whereabouts of people, but some networks are better represented in a *virtual space* (like online) with abstract metrics (like similarity in opinion or interest) defining alternative distance measures between egos.

One of the most important questions in social networks concerns social tie creation. The propensity of two people to become acquaintances or friends may depend on several factors such as co-habitation, common interest, socioeconomic status, age, gender, education level, or occupation, etc. It has been argued in social theory that people who are more similar to each other tend to create new relationships with a higher probability [262]. This selection mechanism [41], called in general *homophily*, has been observed in several studies [77, 26, 237] and has far-reaching consequences in the emergent network structure. The effects of homophily, however, are hardly distinguishable from the consequences of another process called *influence*. This is a force mechanism where connected people become more similar to each other via social or interpersonal influence [41]. This mechanism is arguably behind several macroscopic social phenomena like the spreading of information and memes [37, 374], innovations and services [198, 26] or the emergence of collective social movements [61, 152]. In some cases it is difficult to decide from simple data-driven observations, whether similar people are connected due to homophilic preferences, or became similar via social influence after building the relationship. This question sets one of the most important challenges of computational social science as, depending on the dominant mechanism, competing hypothesis has been set to explain social diffusion processes [26] (as we will discuss later in Section 4.3.1 and 4.5).

Along homophilic preferences, a mechanism called *triadic closure* has been proposed [322, 307] to be a determinant force behind link creation and the emergence of *community structure* in social networks. Triadic closure is a property among three individuals such that if we observe already two ties between them, with a high probability there is or going to be a third tie closing the open triad to become a triangle. This property induces strongly connected local structures built up from several triangles among people usually sharing some common property. Communities in general are connected subgraphs of a network with nodes, which are denser connected with other nodes of their own community than to the rest of the network. Communities can be mutually exclusive partitions, or overlapping sets of nodes, they can be static or dynamic, uni-layer or multilayer, etc. Their detection is an ongoing challenge with several methods and models proposed [126].

The *strength of a social tie* is a linear combination of the amount of time, emotional intensity, intimacy, and reciprocal services, which characterise a social relationship. The strength of social ties are commonly assumed whether being weak or strong and to be strongly correlated with the network position of the actual tie, as argued by Granovetter [155, 153]. In his seminal paper [155] he

suggests that *weak ties* are maintained via sparse interactions, bridging between tightly connected communities keeping the network connected [288], and play important roles in disseminating information globally [78, 76]. On the other hand *strong ties*, sustained by frequent communications, are crucial in shaping the local connectivity of social networks. Due to triadic closure mechanisms they are responsible for the emerging clustered topology [222, 228], and they exert to keep information locally [195, 268].

However, even differentiating ties by strength one cannot simultaneously maintain a large number of social relationships in a meaningful way. This has been argued by Dunbar who proposed the *social brain hypothesis* based on an anthropological correlation between the social group size and neocortex size of primates [110]. He argues that due to cognitive limitations of people, the size of the egocentric network of an individual is also limited to be around 150. Although one could assume that modern communication services and social platforms have helped to overcome this constrain called the *Dunbar's Number*, recent studies have observed a similar limit in number of online friends, which after one cannot commit equally to every acquaintances but needs to share attention [147]. Dunbar further developed his theory by introducing *intimacy circles* by categorising acquaintances by their social tie strength, intimacy and importance to obtain a finer grained structure of one's egocentric network [111, 291].

1.3 Datasets

The central block of my research concerns the collection, curation and analysis of longitudinal datasets recording the actions and interactions of large number of individuals. Most of these datasets are temporally detailed, coupled with geographical informations and demographic details about the anonymised participants. In the following Section I give short descriptions of some important large datasets, which I used recurrently in several studies to make various observations, which will be discussed later during the course of the thesis. Note that all unique IDs in any of the following datasets were anonymised in advance by the providers without the involvement of the researchers, thus individual re-identification of any participants were impossible from the data.

1.3.1 Data representation

Datasets on *individual action dynamics* usually come in a format of action sequences represented in general as a set of events $AC \subset T \times V(\times \prod_i A_i^{ac} \times LOC)$ where T assigns the set of discrete time stamps bounded by the observed period and V is the set of individuals. Usually this tuple is extended by A_i^{ac} sets of different attributes $a_i \in A_i^{ac}$ (like type of action, duration, cost, etc.) of the actual activity, and some LOC location information. This way, e.g., one activity record $ac(t, v, a_1, a_2, \ldots, loc_v)$ indicates that at time $t \in T$ the user $v \in V$ were active in doing $a_1 \in A_1^{ac}$ for $a_2 \in A_2^{ac}$ duration at $loc_v \in LOC$. Such datasets commonly arrive with some accompanying meta-data sets in a general form of $A_j^v \subset \prod_j A_j^v$ describing the attributes j of each user in $v \in V$ regarding different features A_j^v, such as age, gender, home location, active period, etc.

Interaction sequences record temporal networks [173, 232] and usually come in an event sequence format defined as $EV \subset T \times V \times V(\times \prod_i A_i^{ev} \times LOC)$. This representation is very similar to action sequences but it describes the interaction between $(u, v) \in V \times V$ ordered (unordered) pairs of entities performing a directed (undirected) interactions. E.g., in case of mobile call communication events, which are typically called Call Detail Records (CDRs), are like $ev(t, v_{caller}, v_{callee}, duration, cost, loc_{caller})$ and come as long sequences describing all interactions of every users of a mobile phone provider over several months with usually one-second resolution. Using an event sequence and the corresponding meta-data set one can actually define a temporal network (see Chapter 3), where directed interactions can be followed as function of time and space. In addition, accompanying meta-data of users can be defined as a matching feature vector, which

can take more complicated format as explained above, or can be even dynamical, as we will see later. This interpretation is useful to follow the recurrent communication patterns of egos, and the formation of their social ties as a function of time, space, and personal attributes.

Note that several datasets, dynamical or not, could be used to map the *static network structure* of social interactions. I commonly used the standard adjacency list representation [278] of such large networks, or streaming algorithms to process edge list representation of relational data.

1.3.2 Datasets in hand

DS1 - Mobile phone communication data

The datasets I investigated to most in several studies [182, 195, 203, 206, 218, 217, 223, 233, 250, 339, 351, 352, 354, 388] collects sequences of CDRs of mobile-phone calls (MPC) and short messages (SMS) of a large set of individuals. These datasets were recorded by a single operator with 20% market share in an undisclosed European country (ethic statement was issued by the Northeastern University Institutional Review Board). I analysed several variants of these datasets, which typically contained ~ 633 million time stamped phone call and ~ 209 million SMS events recorded during 182 days with 1 second resolution between $6,2$ million ($4,8$ million for SMS) individuals who were connected via $16,8$ million ($10,3$ million for SMS) social ties. In order to take into account only true social interactions and avoid commercial communication, we used only actions which were executed on links between users, who are at least once mutually connected each other.

DS2 - Mobile communication/Credit record data

Another mobile phone data I extensively studied over the last years [246, 238, 237, 239] record the temporal sequence of ~ 8 billion call and SMS CDRs between ~ 112 million anonymised mobile phone users for 21 months in a Latin American country. The dataset was collected by a single operator but other mobile phone users called by the customers of the provider, who were not clients of the actual provider, also appear in the dataset with unique anonymised IDs. The initially constructed static social network contained all users (whether clients or not of the actual provider), while links were drawn between them if they interacted (via call or SMS) at least once during the observation period. In order to filter out call services and other non-human actors from the network, we recursively removed all nodes (and connected links) who appeared with either in-degree $k_{in} = 0$ or out-degree $k_{out} = 0$. We repeated this procedure recursively until we received a network where each user had $k_{in}, k_{out} > 0$, i.e. made at least one outgoing and received at least one incoming communication events during the nearly 2 years of observation. After construction and filtering the network remained with ~ 82 million users connected by ~ 1 billion links.

This mobile call dataset was coupled with another data provided by a Bank from the same country. This data records financial details of ~ 6 million people assigned with unique anonymised identifiers over 8 months. Data records the time varying customer variables as the amount of debit card purchases, purchase categories of purchased goods, monthly personal loans, and static user attributes such as billing postal codes (zip code), the age and gender of customers. A subset of IDs of the anonymised bank and mobile phone customers were matched[5]. This way of combining the datasets allowed us to simultaneously observe the social structure and estimate economic status of the connected individuals. The combined dataset contained ~ 1 million people, all of them assigned with communication events and detailed bank records, connected by ~ 2 million links into a single giant component.

[5]Note, that the matching, data hashing, and anonymisation procedure was carried out through direct communication between the two providers (bank and mobile provider) without the involvement of the scientific partner.

DS3 - Twitter data

In a project we collected a large data corpus from the online news and social networking service, Twitter [174, 15, 245]. There, users can post and interact with messages, "tweets", restricted to 140 characters. Tweets may come with several types of metadata including information about the author's profile, the detected language, where and when the tweet was posted, mentions of other users (denoted by the `@` symbol) and hashtags (denoted by the # symbol) to assign topics, etc. Specifically, we recorded 170 million tweets written in French, posted by 2.5 million users in the timezones GMT and GMT+1 over three years (between July 2014 to May 2017). These tweets were obtained via the Twitter powertrack API feeds provided by Datasift and Gnip with an access rate varying between $15-25\%$. We used this dataset to obtain personal *linguistic data* from the written text of each post and to infer the *social network structure* between users. Tweet messages may be direct interactions between users via mentions by using the symbol (`@username`). We took direct mutual mentions as proxies of social interactions and used them to identify social ties between pairs of users. This constraint lead to a structure of $500,000$ users and 2 million undirected links. In addition, about 2% of tweets contained some *location information* regarding either the tweet author's self-provided position or the place from which the tweet was posted. Note that in some other studies we collected another French Twitter dataset to observe social spreading phenomena [103], while in some other projects [342, 352] we reported results based on the analysis of another, yet very similar, Twitter dataset.

DS4 - Skype service adoption data

In a couple of projects I had the chance to access and analyse the social network of one of the largest online communication services at the time, the Skype network. The centrepiece of this dataset is the contact network, where nodes represent users and edges assign confirmed Skype connections. A user's contact list is composed of friends of mutually confirmed relationship assigned with two time stamps indicating the moment the contact request was approved, and deleted if it happened before the end of the observation period. In some studies we considered the temporal evolution of this network [212], while in others [198, 199] we took its static representation, aggregated for 99 months between September 2003 and November 2011. The largest connected component of this structure includes roughly 510 million users and 4.4 billion edges. In addition, this dataset comes with time stamped purchase records of different Skype services. To study social contagion processes, we followed how users purchase credits for calling phones. For each user, the dataset includes the date when a user first adopted the paid product *buy credit* (first credit purchase, for all purposes). We selected this service since its lifetime of 89 months is considerably long (it was introduced in 2004), and it can be adopted by registered Skype users only. This way the aggregated Skype network provided a complete description of the mediating social structure, which allowed us to calculate the correct network and dynamical properties. To make additional observations, we performed calculations on a second paid service called *subscription*, which was introduced in April 2008, lasts for over 42 months, and could also be adopted by registered Skype users only.

Other datasets

I addition there are several other datasets, which I analysed during my research to make data-driven observations, design data-driven models or to train and verify model predictions. I used various social interaction datasets recording emails [195, 202], Facebook wall posts [269], face-to-face interactions [388], or sexual contacts [217]; individual tracking data of communication, mobility and service usage records [195, 181]; collaboration datasets as scientific co-publication [351, 352] or co-citations [180], business alliances [344], or the election of wikipedia editors [354]; transportation datasets as air-traffic networks [217, 180] or public transportation systems in large cities [20]; or even activity dynamics of firing neurons or earthquakes [202]. In addition, I conducted studies where the focus was to collect a specific dataset in some special settings like face-to-face interactions

in pre-schools, or age contact matrices of people during the COVID pandemic. These data collection efforts will be summarised later together with the results of their analysis. All these datasets provided the solid "observational" ground for my studies, which will be summarised in the following three Chapters.

2. Bursty Human Dynamics

2.1 Introduction

Bursty behaviour is a temporal character of some dynamical systems, which alternate between active periods with high frequency of events and long periods of inactivity. Dynamics characterised by such large temporal fluctuations cannot be explained by the conventional picture using Poisson processes assuming a single temporal scale, but rather can be the result of non-Poissonian dynamics inducing strong temporal heterogeneities on various temporal scales[1].

There are a number of systems in Nature that evolve following non-Poissonian dynamics [200]. One example are earthquakes [92, 100, 36, 101, 326], in which the times of shocks occurring at a given location show bursty temporal patterns, as illustrated in Fig. 2.1(a). Another example are solar flares with bursty emergence induced by huge and rapid releases of energy [261, 375]. It has been shown that the stochastic processes underlying these apparently different phenomena show such universal properties that lead to the same distributions of event sizes, inter-event times, and temporal clustering [101], which can be arguably modelled by the frame of self-organised criticality (SOC) [35]. Burstiness is also seen in the contexts of neuronal systems where the firing of a single neuron (as shown in Fig. 2.1(b)) or collective of neurons appears with such dynamics [211]. Moreover, bursty patterns has been observed in ecology and animal dynamics in the context of initiating conflicts [305], communication, foraging [328], predators waiting in ambush [371], or the displacement of monkeys or mice [64, 275], which form complex self-similar temporal patterns reproduced on multiple time scales very similarly to examples observed in human behaviour. In addition, scale-invariant bursty temporal patterns have also been found in several man-made systems, like written text, in which successive occurrences of the same word display bursty patterns [22]. In case of engineering systems perhaps some of the best examples are in the context of package-based traffic and wireless communication signals, which were found to

[1] Non-Poissonian bursty dynamics is in general characterised by the heterogeneous distribution of inter-event times passing between the consecutive occurrences of a given type of event. In contrast, in a system with Poissonian dynamics, inter-event times are distributed exponentially. However, many empirical inter-event time distributions are broad and follow a log-normal, Weibull, or power-law form, implying that the underlying mechanisms behind them maybe different than a Poisson process.

evolve through non-Poissonian dynamics [80, 179, 236, 296]. Finally, financial markets, in which non-Poissonian dynamics characterises time series of returns of financial assets, stock sales, order books, and other transactions with dynamics studied in the realm of econophysics [256].

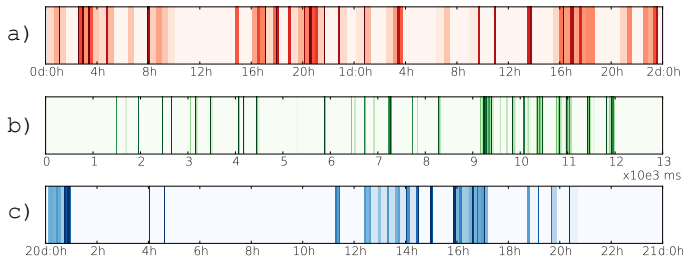


Figure 2.1: (a) Sequence of earthquakes with magnitude larger than two at a single location (south of Chishima Island, 8th–9th October 1994). (b) Firing sequence of a single neuron from a rat's hippocampal. (c) Outgoing mobile phone call sequence of an individual. The shorter the time between the consecutive events are, the darker colour is coded. This figure was published in [202].

Bursty patterns have been found to characterise human dynamics in the timings of actions, dyadic social interactions, or even in collective social phenomena. The first observations were reported by Eckmann *et al.* [116] and by Barabási [39], who observed broad inter-event time distributions with a power-law tail by analysing datasets of email correspondence. These seminal papers initiated an avalanche of studies to observe, characterise, and model bursty phenomena detected in a number of human activities. Various examples were found, like emails [116, 39], letter correspondence [287], mobile phone calls and short messages [202] (like in Fig. 2.1(c)), web browsing [105], printing [161], library loans [361], job submission to computers [219], and file transfer in computer network [296], or even in arm movements of human subjects [87], just to mention a few. In addition, further examples were identified at the group or societal level, such as the emergence of causal temporal motifs [224], the evolution of mass demonstrations, revolutions, global information cascades, and wars [62, 338].

All these new observations highlighted some shortcomings of earlier methods to characterise human bursty dynamics and called for novel measures and models to gain deeper understanding about the roots of bursty patterns in human behaviour. Several modelling frameworks of bursty human dynamics have been proposed over the last years, which could be roughly classified into three main groups based on the assumed underlying explanatory mechanisms. In his original study, Barabási suggested that bursty activity patterns could be the consequence [39, 287, 361] that people do not execute their "to-dos" in a random fashion but assign importance to each task at hand. This induces intrinsic correlations between different tasks and results in bursty patterns of completed activities, which can be effectively modelled by *priority queues* with different constraints. Another direction was proposed by Malmgren *et al.* [255, 254] who argued that "human behaviour is primarily driven by external factors such as circadian and weekly cycles, which introduces a set of distinct characteristic time scales, thereby giving rise to heavy tails". This approach assumes no intrinsic correlations in human activities but models the dynamics as alternating homogeneous and non-homogeneous *Poisson processes*. The third main modelling concept assumes strong correlations between consecutive events and define *non-Markovian dynamics* by using memory functions [360, 160], self-exciting point processes [259, 184], or reinforcement mechanisms [203, 365] in simulating bursty activity patterns. Finally, several other modelling ideas were suggested assuming self-organised criticality [338], local structural correlations [274], some dynamical process like random walk [144], contact process in extended Griffiths phases [286], or voter model [124] to introduce heterogeneous temporal patterns at the individual or system levels.

Based on these advancements more far-reaching scientific questions have been addressed about

the effects of non-Poissonian patterns of individuals on collective dynamical processes. First question raised, whether they are ongoing or co-evolving with bursty actions and interactions. An important example is diffusion processes on temporal networks where bursty dyadic interactions may enhance or slow down the speed and/or control the emergence of globally spreading process, like information diffusion, epidemics, or random walk [173]. Beyond the conventional modelling and simulation techniques of such processes, data-driven models and random reference systems [195, 268] were recently shown to be very successful in addressing such problematics.

I entered this field during my first postdoctoral period at Aalto University and worked with several colleagues on various topics to observe, measure, and model bursty human behaviour, and to better understand its consequences on the evolution of dynamical processes. On the methodological level I had two main contributions: I proposed an entirely new measure to detect bursty temporal correlations in heterogeneous signals [202] and, together with colleagues, a method to account for the effects of circadian fluctuations to identify to what extent they are responsible for the emergence of bursty patterns in human dynamics [182]. Using these techniques I conducted data-analysis studies [212, 203] to observe and characterise bursty social link creation and maintenance, or the bursty dynamics of recurrent higher order patterns in temporal networks [75]. I also developed various models using reinforcement mechanisms to explain individual and dyadic correlated bursty behaviour [202, 203]. We were among the firsts to use random reference models to identify the effects of burstiness on spreading processes [195, 218] and to develop a generative temporal network model with bursty interactions [352]. In addition, together with collaborators I recently published a monograph book [200] to review the knowledge accumulated during the last ten years in this domain.

In this Chapter, without aiming a complete overview of the field, I give a brief summary of my main contributions to the area. After this introduction, I lay down some general concepts and measures which I will rely on later during the Chapter. Then, I organise my description by first introducing my methodological contributions, then observational studies, and modelling, and finally I will conclude my understanding and contribution to the overall scientific landscape.

2.2 Characterisation of bursty phenomena

Dynamical systems can be described as time series of sequential observations [63] where timing of an observation, denoted by t, can be either continuous or discrete. Since most datasets of human dynamics we analyse have been recorded digitally, we will here focus on the case of discrete timings. In this sense, the time series can be called an event sequence, where each event indicates an observation with a particular character. In this series the ith event takes place at time t_i with the result of the observation z_i describing the actual state of the system with a number, set of numbers, etc[2]. At the simplest scenario, we assume that the system at a given time can be in two states only, as being active and performing an event, or being inactive. The event sequence with n events can be represented by an ordered list of event timings, i.e., $ev(t_i) = \{t_0, t_1, \cdots, t_{n-1}\}$, where t_i denotes the timing of the ith event. Such dynamics (also called point processes) can be described in a form of binary event sequences of $x(t)$ that takes a value of 1 at time $t = t_i$ of events, or 0 otherwise. Formally, for discrete timings, one can write the signal as $x(t) = \sum_{i=0}^{n-1} \delta_{t,t_i}$, where δ denotes the Kronecker delta.

The Poisson process

The temporal Poisson process is a stochastic process, which is commonly used to model random processes such as photon counting in astronomical observations [318], the arrival of customers at a store, or packages at a router, just to mention a few examples. It evolves via completely

[2]Note that some events could occur in a time interval or with duration, like phone calls between individuals [173], what we neglect in our description at the outset unless stated otherwise.

independent events, thus it can be interpreted as a type of continuous-time Markov process. In a Poisson process, the probability that n events occur within a bounded interval follows a Poisson distribution $P(n) = \frac{\lambda^n e^{-\lambda}}{n!}$, where λ denotes the average number of events per interval, which is equal to the variance of the distribution in this case. Since these stochastic processes consist of completely independent events, they have served as reference models when studying bursty systems. As we will see later, bursty temporal sequences emerge with fundamentally different dynamics with strong temporal heterogeneities and temporal correlations. Any deviation in their dynamics from the corresponding Poisson model can help us to indicate patterns induced by correlations or other factors like memory effects.

Throughout the thesis we are going to refer to two types of Poisson processes. One is called the *homogeneous Poisson process*, which is characterised by a constant event rate λ, while the other type, called the *non-homogeneous Poisson process*, defined such that the event rate varies over time, denoted by $\lambda(t)$. For more precise definitions and discussion on the characters of Poisson processes we suggest the reader to study the extended literature addressing this process, e.g., Ref. [156].

The inter-event time and residual time

The first and most important measure to characterise bursty temporal sequences is based on the quantity called the *inter-event time*, $\tau_i \equiv t_i - t_{i-1}$, defined as the time interval between two consecutive events at times t_{i-1} and t_i for $i = 1, \cdots, n-1$. For an event sequence of $n \geq 2$, we can obtain the sequence of inter-event times, i.e., $iet(\tau_i) = \{\tau_1, \cdots, \tau_{n-1}\}$, and compute their probability density function, i.e., the *inter-event time distribution* $P(\tau)$. For completely regular time series, all inter-event times are the same and equal to the mean inter-event time, denoted by $\langle\tau\rangle$, thus the inter-event time distribution appears as:

$$P(\tau) = \delta(\tau - \langle\tau\rangle), \tag{2.1}$$

where $\delta(\cdot)$ denotes the Dirac delta function. Here the standard deviation of inter-event times, denoted by σ, is zero.

For the completely random and homogeneous Poisson process, it is easy to derive [156] that the inter-event times are exponentially distributed as follows:

$$P(\tau) = \frac{1}{\langle\tau\rangle} e^{-\tau/\langle\tau\rangle}, \tag{2.2}$$

where $\sigma = \langle\tau\rangle$ and the event rate is $\lambda = 1/\langle\tau\rangle$.

In many empirical processes inter-event time distributions have been observed to be broad with heavy tails ranging over several magnitudes. In such bursty time series the fluctuations characterised by σ are much larger than $\langle\tau\rangle$, indicating that $P(\tau)$ is rather different from an exponential distribution, as it would derive from Poisson dynamics. Bursty systems evolve through events that are heterogeneously distributed in time and exhibit a broad $P(\tau)$, which can be fitted with either power law, log-normal, or stretched exponential distributions, just to name a few candidates. Most commonly, they can be approximated by a power-law distribution function with an exponential cutoff, defined as

$$P(\tau) \simeq C\tau^{-\alpha} e^{-\tau/\tau_c}, \tag{2.3}$$

where C denotes a normalisation constant, α is the power-law exponent, and τ_c sets the position of the exponential cutoff. The power-law scaling of $P(\tau)$ indicates the lack of any characteristic time scale, but the presence of strong temporal fluctuations, characterised by the power-law exponent α. Power-law distributions are also associated to the concepts of scale-invariance and self-similarity [277] and deemed to have an important meaning, especially in terms of universality classes in statistical physics [301].

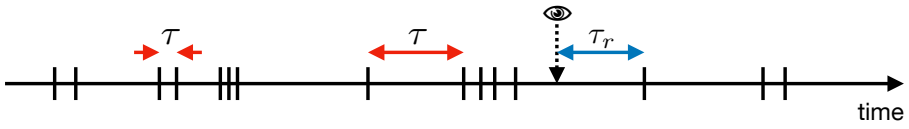


Figure 2.2: Schematic diagram of an event sequence, where each vertical line indicates the timing of an event. The inter-event time τ is the time interval between two consecutive events. The residual time τ_r is the time interval from a random moment (e.g., the timing annotated by the vertical arrow) to the next event. This figure was published in [200].

Note that there is another similar quantity, called the *residual time* τ_r (also called the residual waiting time), what we will use later during our discussion. It's definition considers that the observations of an event sequence always cover a finite period and usually begins at a random moment of time. The time interval between the time of the observation and the first observed event is the residual time τ_r. Its distribution and average can be derived from the corresponding inter-event time distribution as

$$P(\tau_r) = \frac{1}{\langle\tau\rangle}\int_{\tau_r}^{\infty} P(\tau)d\tau, \qquad \langle\tau_r\rangle = \int_0^{\infty} \tau_r P(\tau_r)d\tau_r = \frac{\langle\tau^2\rangle}{2\langle\tau\rangle}. \tag{2.4}$$

This result explains a phenomenon called the *waiting-time paradox*, which has important consequences in dynamical processes evolving on bursty temporal systems as will be discussed later.

The burstiness parameter

The heterogeneity of the inter-event times can be quantified by a single measure introduced by Goh and Barabási [146]. The burstiness parameter B is defined as the function of the coefficient of variation (CV) of inter-event times $r \equiv \sigma/\langle\tau\rangle$ to measures temporal heterogeneity as follows:

$$B \equiv \frac{r-1}{r+1} = \frac{\sigma - \langle\tau\rangle}{\sigma + \langle\tau\rangle}. \tag{2.5}$$

Here B takes the value of -1 for regular time series with $\sigma = 0$, and it is equal to 0 for random, Poissonian time series where $\sigma = \langle\tau\rangle$. In case when the time series appears with more heterogeneous inter-event times than a Poisson process, the burstiness parameter is positive ($B > 0$), while taking the value of 1 only for extremely bursty cases with $\sigma \to \infty$. Note that this measure has been recently shown to have some finite size effect, and an alternative measure has been introduced to account for these shortcomings [213].

The autocorrelation function

The conventional way for detecting correlations in time series is to measure the autocorrelation function. To define we use the representation of event sequences as binary signals $x(t)$ and introduce the delay time t_d, which sets a time lag between two observations of the signal $x(t)$. Then the autocorrelation function with delay time t_d is defined as follows:

$$A(t_d) \equiv \frac{\langle x(t)x(t+t_d)\rangle_t - \langle x(t)\rangle_t^2}{\langle x(t)^2\rangle_t - \langle x(t)\rangle_t^2}, \tag{2.6}$$

where $\langle\cdot\rangle_t$ denotes the time average over the observation period [63]. In time series with temporal correlations, $A(t_d)$ typically decays as a power law:

$$A(t_d) \sim t_d^{-\gamma} \tag{2.7}$$

with decaying exponent γ. In addition, this measure can be related to the power spectrum or spectral density of the signal $x(t)$ as follows:

$$P(\omega) = \left|\int x(t)e^{i\omega t}dt\right|^2 = \int A(t_d)e^{-i\omega t_d}dt_d, \tag{2.8}$$

which appears as the Fourier transform of autocorrelation function, indicating dominant ω event frequencies present in the signal.

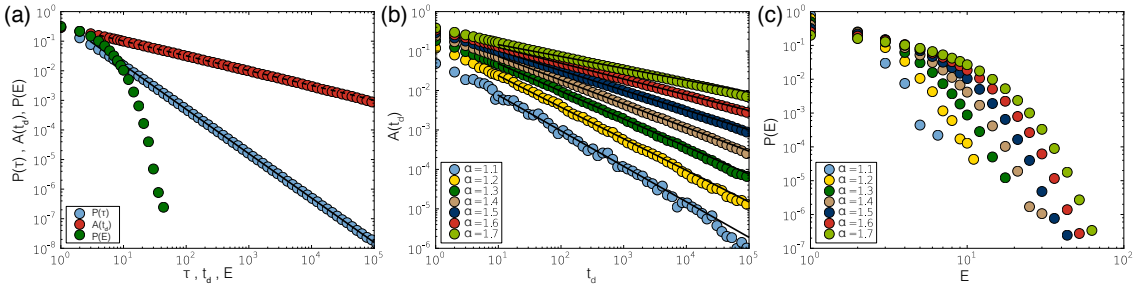


Figure 2.3: The characteristic functions calculated for heterogeneous independent signals. (a) $P(\tau)$, $A(t_d)$ and $P(E)$ functions for $\alpha = 1.5$. Solid line is a power-law function with the given α exponent value, while dashed line denotes a a power-law function with an effective 0.5 exponent value. (b) $A(t_d)$ effective autocorrelation functions for various α exponents. Straight lines are denoting power-law functions with α exponents satisfying the $\alpha + \gamma = 2$ relation. (c) Corresponding $P(E)$ distributions for various α exponents. This figure was published in [202].

A scaling relation between the α inter-event time and the γ autocorrelation exponents has been studied both analytically and numerically [251, 21, 357]. It has been shown that they relate as

$$\begin{aligned} \alpha + \gamma &= 2 \quad \text{for } 1 < \alpha \leq 2, \\ \alpha - \gamma &= 2 \quad \text{for } 2 < \alpha \leq 3. \end{aligned} \tag{2.9}$$

This indicates that the power-law decaying autocorrelation function could be explained solely by the inhomogeneous inter-event times and not by temporal correlations in the time series. In fact, the observed autocorrelation functions measure not only correlations between events but also between consecutive inter-event times of arbitrary length. Such correlations spuriously appear in independent heterogeneous time series disallowing autocorrelation to be a proper measure of real temporal correlations in bursty sequences. This is demonstrated in Fig. 2.3(a) where we built an independent event sequence by sampling randomly inter-event times from a $P(\tau) \sim \tau^{-\alpha}$ power-law distribution with exponent $\alpha = 1.5$ (blue symbols) and measured the $A(t_d)$ autocorrelation function in this uncorrelated signal. Due to the heterogeneity of the inter-event time distribution effective positive correlations are indicated by the autocorrelation, which emerges with a power-law tail (red symbols) even the sequence is independent. Fig. 2.3(b) demonstrates that the scaling exponent γ of the emerging autocorrelation function is dependent on the α inter-event time distribution, in full agreement with the relation suggested in Eq. 2.9.

2.2.1 The Bursty Train Size Distribution

This ambiguity to detect short term temporal correlations in bursty event sequences motivated us to provide a new measure, which can decide evidently the presence of such dependencies, independently from the shape of the inter-event time distribution. More precisely, we made the qualitative observation that bursty events do not usually come in pairs but may form longer trains, where consecutive events may be in a causal relation with each other. However, to detect such bursty clusters in binary event sequence, $x(t)$, first we have to identify those events which we consider to be causally correlated. The smallest temporal scale at which correlations can emerge in the dynamics is between consecutive events. If only $x(t)$ is known, we can assume two consecutive actions at t_i and t_{i+1} to be related if they follow each other within a short time interval, $t_{i+1} - t_i \leq \Delta t$ [380, 349]. For events with the duration d_i this condition is slightly modified: $t_{i+1} - (t_i + d_i) \leq \Delta t$.

This definition allows us to detect bursty periods, defined as a sequence of events where each event follows the previous one within a time interval Δt, as illustrated in Fig. 2.4. By counting

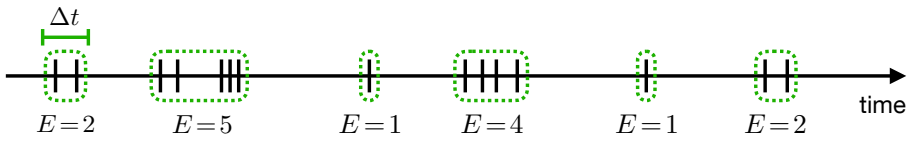


Figure 2.4: Schematic diagram of an event sequence, where each vertical line indicates the timing of the event. For a given time window Δt, a bursty train is determined by a set of events separated by $\tau \leq \Delta t$, while events in different trains are separated by $\tau > \Delta t$. The number of events in each bursty train, i.e., bursty train size, is denoted by E. This figure was published in [200].

the number of events, E, that belong to the same bursty period, we can calculate their distribution $P_{\Delta t}(E)$. For a sequence of independent events, $P_{\Delta t}(E)$ is uniquely determined by Δt and the inter-event time distribution $P(\tau)$ as follows:

$$P_{\Delta t}(E) = \left(\int_0^{\Delta t} P(\tau)d\tau\right)^{E-1}\left(1-\int_0^{\Delta t} P(\tau)d\tau\right) \approx \frac{1}{E_c(\Delta t)}e^{-E/E_c(\Delta t)} \tag{2.10}$$

for $E > 0$. Here the integral $F(\Delta t) = \int_0^{\Delta t} P(\tau)d\tau$ defines the probability to draw an inter-event time $P(\tau) \leq \Delta t$ randomly from an arbitrary distribution $P(\tau)$, $E_c(\Delta t) \equiv \frac{1}{-\ln F(\Delta t)}$ is a constant, and approximation appears due to the series expansion of the constant multiplicative term. The first term on the l.h.s. of Eq.2.10 gives the probability that we draw an inter-event times $P(\tau) < \Delta t$ independently $E-1$ consecutive times, while the second term assigns that the E^{th} drawing gives a $P(\tau) > \Delta t$ therefore the evolving train size becomes exactly E. If the measured time window is finite ($\Delta t < \infty$), which is always the case here, the integral $\int_0^{\Delta t} P(\tau)d\tau < 1$ is a constant and the asymptotic behaviour appears like in a general exponential form. Consequently, for any finite independent event sequence the $P_{\Delta t}(E)$ distribution decays exponentially even if the inter-event time distribution is fat-tailed. Deviations from this exponential behaviour indicate correlations in the timing of the consecutive events.

Bursty sequences in human communication

To check the scaling behaviour of $P_{\Delta t}(E)$ in real systems we focused on outgoing events of individuals in three selected communication datasets: (a) A mobile-call dataset from a European operator (see DS1 in Section 1.3.2); (b) Text message records from the same dataset (also DS1); and (c) Email communication sequences [116]. For each of these event sequences, the distribution of inter-event times measured between outgoing events are shown in Fig.2.5 (left bottom panels) and the estimated power-law exponent values are summarised in Table 2.1. The autocorrelation functions, which were averaged over $1,000$ randomly selected users with maximum time lag of $t_d = 10^6$, indicate strong temporal correlation (as seen in Fig.2.5.a and b (right bottom panels) with exponents in Table 2.1). The power-law behaviour in $A(t_d)$ appears after a short period, denoting the reaction time through the corresponding channel, and lasts up to 12 hours, capturing the natural rhythm of human activities. For emails in Fig.2.5.c (right bottom panels) long term correlation are detected up to 8 hours, which reflects the typical length of office hours (note that the dataset includes internal email communication of a university staff).

The broad shape of $P(\tau)$ and $A(t_d)$ functions confirm that human communication dynamics is heterogeneous and displays non-trivial correlations up to finite time-scales. However, after destroying event correlations by shuffling inter-event times in the interaction sequence of single individuals (see a how to in Section 3.3) the autocorrelation function still shows slow power-law like decay (empty symbols on bottom right panels) via spurious residual dependencies. This clearly demonstrates the disability of $A(\tau)$ to characterise correlations for heterogeneous signals, just as we have already seen for modelled signals. However as we have discussed earlier, the $P_{\Delta t}(E)$

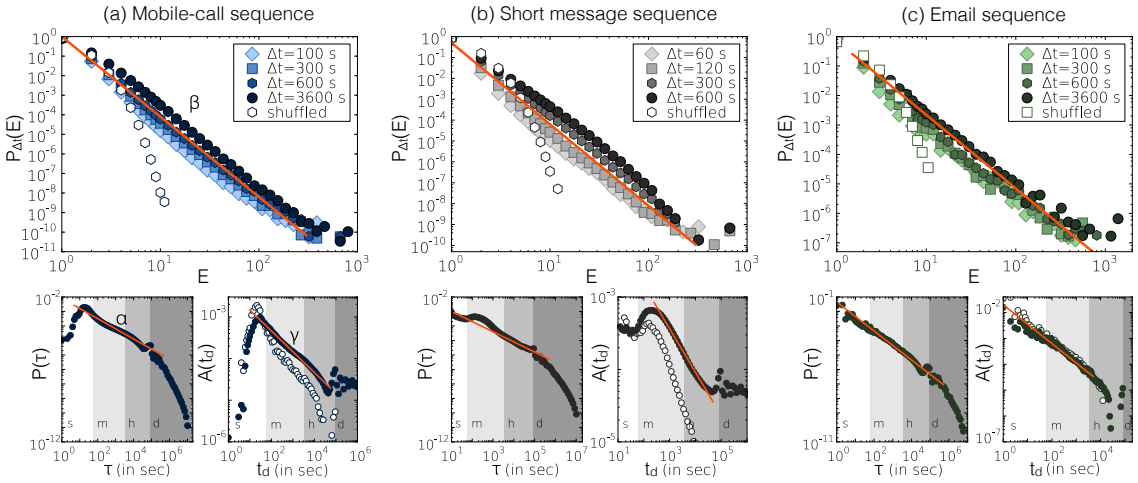


Figure 2.5: The bursty train size distribution $P_{\Delta t}(E)$ with various time windows Δt (main panels), the inter-event time distribution $P(\tau)$ (left bottom panels), and autocorrelation functions $A(t_d)$ (right bottom panels) for different human communication datasets such as (a) Mobile phone call dataset: The scale-invariant behaviour was characterised by power-law functions with exponent values $\gamma \simeq 0.5$, $\beta \simeq 4.1$, and $\alpha \simeq 0.7$ (b) Short message sequences taking values $\gamma \simeq 0.6$, $\beta \simeq 3.9$ and $\alpha \simeq 0.7$. (c) Email event sequence with exponents $\gamma \simeq 0.75$, $\beta \simeq 2.5$ and $\alpha = 1.0$. A gap in the tail of $A(t_d)$ on figure (c) appears due to logarithmic binning and slightly negative correlation values. Empty symbols assign the corresponding calculation results on independent sequences. Vertical stripes called s, m, h and d are denoting seconds, minutes, hours and days, respectively. This figure was published in [202].

	α	β	γ	ν
Mobile-call sequence	0.7	4.1	0.5	3.0
Short message sequence	0.7	3.9	0.6	2.8
Email sequence	1.0	2.5	0.75	1.3
Model	1.3	3.0	0.7	2.0

Table 2.1: Characteristic exponents of the (α) inter-event time distribution, (β) bursty train size, (γ) autocorrelation functions and ν memory functions calculated in different datasets and for the model study (see Section 1.3.2). This table was published in [200]

distribution should indicate evidently the presence of short temporal correlations. Calculating this distribution for various Δt windows, we find that it depicts a scale invariant behaviour as

$$P(E) \sim E^{-\beta} \tag{2.11}$$

for each of the empirical event sequences as shown in the main panels of Fig.2.5. $P_{\Delta t}(E)$ evidently indicates that there are strong temporal correlations in the empirical sequences as it is remarkably different from the corresponding distributions calculated for independent sequences which, as predicted by (2.10), appear with exponential decay (empty symbols on the main panels).

Exponential behaviour of $P_{\Delta t}(E)$ was also expected from results published in the literature assuming human communication behaviour to be uncorrelated [255, 380, 24]. However, the observed scaling behaviour of $P_{\Delta t}(E)$ offers a direct evidence of correlations in human dynamics, which can arguably be responsible for the observed bursty dynamics. These correlations induce long bursty trains in the event sequence rather than short bursts of independent events. In addition, we have found that the scaling of the $P_{\Delta t}(E)$ distribution is quite robust against the choice of Δt for an extended regime of time-window sizes, or when it is computed for individuals or group people of

similar activity level, or once the effects daily fluctuations are accounted for (for results see [202]).

Interestingly, in a subsequent study [75] we found similar correlated bursty trains of higher order structures in temporal networks. Patterns like k-order hyperedges were emerging in long correlated bursty trains in consecutive aggregation time windows recording face-to-face interactions of people at different settings (school, workplace, hospital, conference etc.). Further we studied the transition of different higher-order structures within bursty trains and observed that transitions among different group sizes are more abrupt in settings delineated by planned and regular activities (like workplace and hospitals) and more smooth in environments characterized by more spontaneous contacts (like university campus or conferences).

Note that we observed long bursty event trains and similar scaling of their size in various natural phenomena, like in earthquake sequences recorded at given locations [326, 390], or in the firing patterns of single neurons recorded in rat's hypocampal. Corresponding results are not shown here but reported in [202].

2.3 Cyclic patterns in human dynamics

It is evident that humans follow intrinsic periodic patterns of circadian, weekly, and even longer cycles [255, 182, 19]. Such cycles clearly contribute to the inhomogeneities of temporal patterns, and they often result in an exponential cutoff to the inter-event time distributions. Identifying and filtering out such cyclic patterns from a time series can reveal bursty behaviour of different origins than those cycles [182]. In order to characterise such cyclic patterns, let us consider a time series, i.e., the number of events at time t, denoted by $x(t)$, for the entire period of $0 \leq t < T$. One may be interested in a specific cycle, like daily or weekly ones, with period denoted by $T_{\circlearrowleft}$. Then, for a given period of $T_{\circlearrowleft}$, the event rate with $0 \leq t < T_{\circlearrowleft}$ can be defined as

$$\rho(t) \equiv \frac{T_{\circlearrowleft}}{X} \sum_{k=0}^{T/T_{\circlearrowleft}} x(t+kT_{\circlearrowleft}), \qquad X \equiv \int_0^T x(t)dt. \tag{2.12}$$

Such cycles turn out to be also apparent in the inter-event time distributions $P(\tau)$. For example, one finds peaks of $P(\tau)$ corresponding to multiples of one day in mobile phone communication as can be seen in Fig. 2.5a and b lower left panels. Note that such periodicities could be characterised by means of a power spectrum analysis in Eq. (2.8), however here we take another way.

Once such cycles are identified in terms of the event rate $\rho(t)$, we can filter them by deseasoning the time series [182]. First, we extend indefinitely the domain of $\rho(t)$ by $\rho(t+kT_{\circlearrowleft}) = \rho(t)$ with an arbitrary integer k. Then using the identity of $\rho(t)dt = \rho^*(t^*)dt^*$ with the deseasoned event rate of $\rho^*(t^*) = 1$, we can get the de-seasoned time $t^*(t)$ as

$$t^*(t) \equiv \int_0^t \rho(t')dt'. \tag{2.13}$$

For the schematic example of the de-seasoning method, see Fig. 2.6a. In plain words, the time is dilated (respectively contracted) at the moment of the high (respectively low) event rate resulting an overally constant average event rate as shown in as demonstrated in Fig. 2.6b. Then the de-seasoned event sequence of $\{t^*(t_i)\}$ is compared to the original event sequence of $\{t_i\}$ to see how strong signature of burstiness or memory effects remained in the de-seasoned sequence. This reveals whether the empirically observed temporal heterogeneities can (or cannot) be explained by the intrinsic cyclic patterns, characterised in terms of the event rate. For example, one can measure the burstiness parameter B_T for both the original and the de-seasoned mobile phone call series as shown in Fig. 2.6c using DS1 (see Section 1.3.2). Although the $P(B_T)$ distribution slightly changes after de-seasoning over a period T, it assigns that the majority of individual sequences remain with positive bursty parameters.

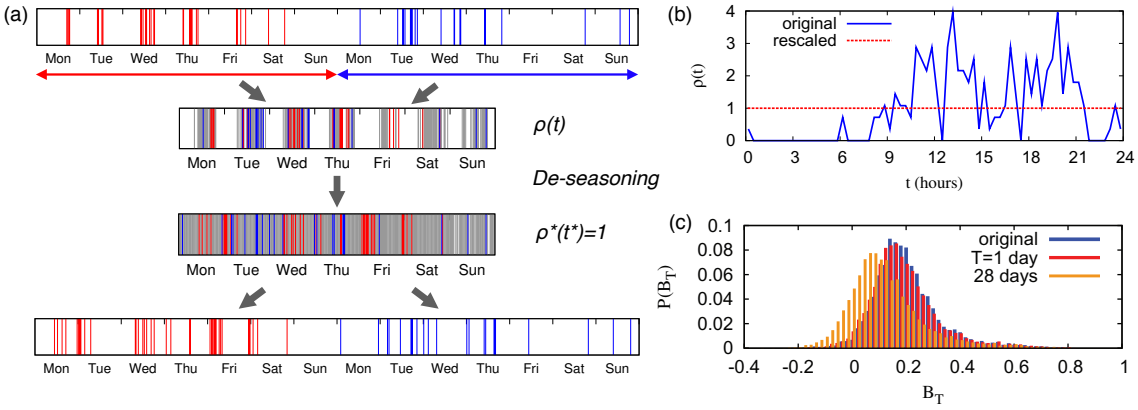


Figure 2.6: (a) An example of the de-seasoning method applied to a mobile call series of a user, with $T_{\circlearrowleft} = 1$ week. The top shows the first two weeks of the call series coloured in red (the first week) and blue (the second week). Events for all weeks are collected in one week period to obtain the event rate $\rho(t)$ for $0 \leq t < T_{\circlearrowleft}$. After de-seasoning, the events in each week are put back to their original slot. (b) The original (blue) and de-seasoned (red) hourly event rate of communication for individuals with 200 calls. (c) The distributions of B_T bursty parameters of individual users with the same strength after de-seasoning over $T_{\circlearrowleft}$ period. This figure was prepared by HH. Jo and published in [182].

One can also obtain the de-seasoned inter-event time τ_i^* corresponding to the original inter-event time $\tau_i = t_i - t_{i-1}$ as

$$\tau_i^* \equiv t^*(t_i) - t^*(t_{i-1}) = \int_{t_{i-1}}^{t_i} \rho(t')dt'. \tag{2.14}$$

This way the de-seasoned inter-event time distribution $P(\tau^*)$ can be compared to the original inter-event time distribution $P(\tau)$. As shown in Fig. 2.7a-c, the inter-event time distributions for the original and de-seasoned event sequences show almost the same shape for various values of $T_{\circlearrowleft}$ and for individuals of various activity level. At the same time, it is evident from Fig. 2.7d-f, that after de-seasoning the circadian and weekly peaks of ω event frequencies disappear from the power spectra (for definition see Eq. 2.8), while its overall scaling remains very similar to the original spectrum. All these results together imply that bursty human dynamics cannot be exclusively explained by periodic circadian and weekly fluctuations, but it may have some other intrinsic behavioural origins.

I have another set of works [195, 218], which provide new tools to understand the effects and consequences of bursty interactions, through the introduction of random reference models of temporal networks. These works will be addressed in Section 3.3 as they provide tools to analyse temporal networks in general, not only bursty phenomena exclusively.

2.4 Observation of bursty phenomena

Bursty dynamics characterise human behaviour on the individual level and in turn may determine the evolution of dyadic interactions or the emergence of macroscopic phenomena in the social network. The dynamics of social networks can be discussed in terms of nodes, links, communities, or at the collective level, and can be characterised at different temporal scale, as we will discuss later in Chapter 3 on temporal networks. To address bursty dynamics of interactions, we distinguish between two temporal scales of link dynamics, which can be assigned to rather different types of behaviour. On one hand, we consider the slow dynamics of social link creation and decay, which determines the evolution of the social network. As an example, think about student mates with whom one may maintain a social relationship over years, which typically decay after graduation.

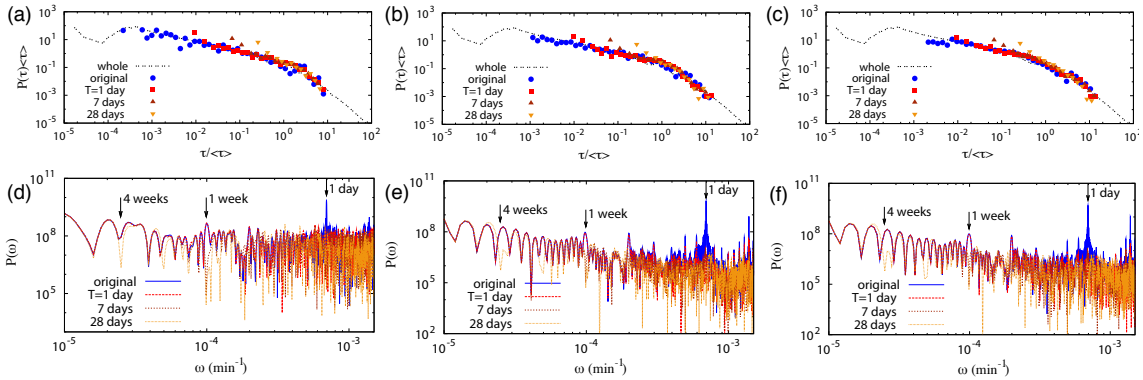


Figure 2.7: (a-c) The original and de-seasoned inter-event times extracted from the individual call sequences of mobile phone users. Dashed line assigns the $P(\tau)$ of the whole population. Rescaling were done for periods of $T = 1$, 7, and 28 days. (e-f) Corresponding power-spectrum curves as the function ω event frequencies with unites $1/min$. Individual users with $s = 400$ (a, d), 800 (b, e), 1600 (c, f) number of calls were analysed. This figure was prepared by HH. Jo and published in [182].

On the other hand, we consider temporal interactions appearing with a rapid pace on existing social ties. These are for example calls, messages, or emails, which typically appear recurrently with high frequency and short duration as compared to the lifetime of a social tie. Next, I summarise some observational results we obtained on real social networks, published in [212, 203], to see whether bursty characters appear in the interaction dynamics at these two temporal scales, and if yes, how these bursty interactions are distributed in the egocentric network of an individual.

2.4.1 Bursty egocentric network evolution

First let's concentrate on the evolution of egocentric networks by analysing creation and decay of confirmed contacts in the online social-communication system of Skype [212]. As we have already explained in Section 1.3.2, the DS4 dataset contains the time stamps of approval of each Skype contacts (which can be regarded as times of link creation), and deletion of Skype contacts if it happened before the end of the observation period. In addition, we consider time series indicating the number of days in each month when the user connected to the Skype network, and also the adoption time (first usage) of each free service [e.g. Skype-to-Skype (S2S) audio calls, video calls, chat, etc.] together with time series indicating the number of days in each month when the user used the given service.

To characterise the temporal evolution of contact lists, we first examine the sequences of edge addition and deletion of each individual by calculating the distributions of inter-event times defined as $\tau_a = t^a_{i+1} - t^a_i$ (resp. $\tau_d = t^d_{i+1} - t^d_i$) between consecutive additions (resp. deletions) events of the same user. As demonstrated in Fig. 2.8a, these distributions are very heterogeneous both in case of edge additions and edge deletions. They show rather similar scaling, which can be approximated with a power-law function with exponent $\gamma \simeq 0.85$ and an exponential cutoff. This is an interesting observation as one would expect rather different decision mechanisms behind adding and deleting a contact as additions can be assumed to be driven by the desire or need to communicate or to signal a social relation, while deletions are arguably driven by the desire not to be visible or accessible by the deleted contact.

To identify temporal correlations between consecutive actions of link additions or deletions, we locate bursty event trains and compute their size distributions using the methodology explained in Section 2.2.1. More precisely, we analysed the edge modification sequences of each individual and extracted the clusters of events of new edge addition and deletion (the trains) to record their size E_a (resp. E_d). The fact that the $P(E_a)$ (resp. $P(E_d)$) distribution in Fig.2.8 spans over orders of

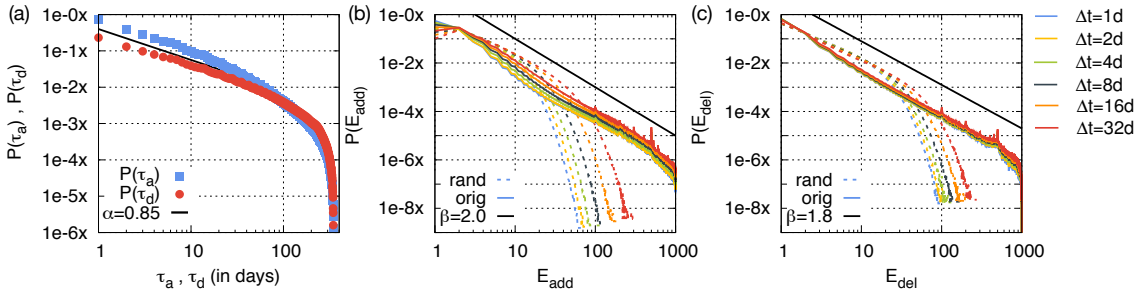


Figure 2.8: (a) Inter-event time distributions of edge addition (blue squares) and deletion (red circles) events of users. The straight line indicates a power-law function with exponent $\alpha = 0.85$. (b,c) Distribution of number of events in bursty trains of (b) contact addition and (c) deletion of individuals. Distributions were calculated with time window sizes $\Delta t = 1$, 2, 4, 8, 16 and 32 days. Distributions for randomly shuffled sequences (dashed line) were calculated with the same Δt values. Straight lines indicate power-law functions with exponents (b) $\beta = 2.0$ and (c) $\beta = 1.8$. This figure was prepared by R. Kikas and published in [212].

magnitude, for several Δt values, confirming the presence of correlations between consecutive events of edge additions (resp. deletions). This is even more apparent once we compare the empirical $P(E)$ functions to the equivalent distributions calculated for independent sequences where inter-event times has been randomly shuffled. It puts into evidence that the actions of an individual are not independent and that the evolution of egocentric networks is not only heterogeneous in time but driven by intrinsic correlations. They lead to the presence of high activity bursty periods, where a large number of edges are added or deleted, separated by long low activity intervals.

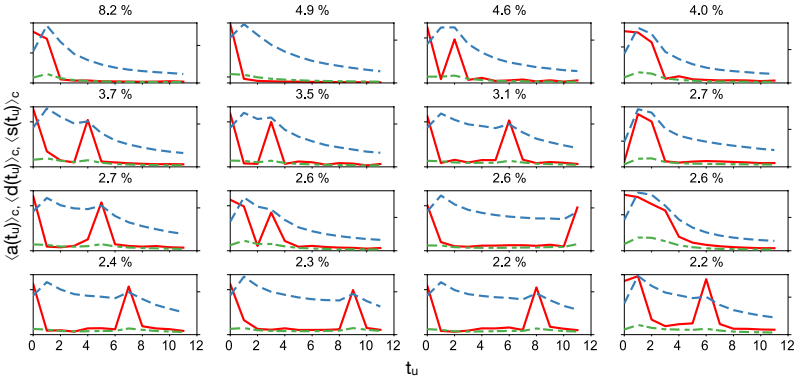


Figure 2.9: Groups of contact addition patterns, where the average number of $\langle a(t_u)\rangle_c$ new contacts added (red), $\langle d(t_u)\rangle_c$ connected days (blue), and $\langle s(t_u)\rangle_c$ days used Skype-to-Skype service at the actual month are shown. Left vertical axe is the added contact number, right axe is in days. This figure was prepared by R. Kikas and published in [212].

So far we have observed that edge addition and deletion events of an individual are bursty and clustered in time, yet we know less about when these bursty trains evolve during the lifetime of a user. Do they appear in any time or there are typical activity patterns of edge additions or maybe triggered by other user actions? To answer these questions, we compare the edge addition sequences of individuals during the first year of their t_u user time after their registration [212]. We keep track the a_i number of newly added edges of each node i in each month to obtain a discrete $a_i(t_u)$ sequence for each individual with $t_u = 1...12$. To be able to compare sequences of users with diverse overall intensity we applied the *Symbolic Aggregate Approximation* (SAX) method [249] with alphabet size 10, while to detect groups of users with similar edge addition dynamics, we applied the *k-means clustering* method on the activity sequences using euclidean distance. We determined the optimal cluster number to be 44 by using the *Elbow method*. For the most populated clusters, the average link addition activity curves, $\langle a_i(t_u)\rangle_k$, are shown in Fig. 2.9 (red line). Looking at the most common patterns it is straightforward that typically people perform their principal (the largest and usually the only one) edge addition burst right after they join the network, a

behaviour which is confirmed by other studies [131]. This is the time when they explore their social acquaintances who have already joined the system, while later they add contacts just occasionally with lower frequency. This behaviour and its significance has been demonstrated in [212] (results not shown here). However, a single bursty peak at early time is not the characteristic of every user. Less common motifs in Fig.2.9 show that principal bursty peaks may emerge later or even in multiple times. This observation indicates that events other than user registration (e.g. adoption of different services) may also trigger immediate changes in the egocentric network. Actually, we found that the link addition dynamics is positively correlated (with coefficient $r = 0.308137$) with the overall activity of users (shown with blue lines in Fig.2.9), what we measured as the average number of connected days per month. Similarly, we found correlations with the activity of free service usage (with coefficient $r = 0.34608$), shown as green lines. More importantly, we showed that the probability of a user performing a link addition burst is strongly conditional to the adoption of free and payed services, which explains the emergent later peaks in the link addition dynamics (for further results see [212]).

2.4.2 Bursty communication in egocentric networks

In our second empirical study let's move from the level of social tie evolution to the level of time varying interactions in egocentric networks. Analysing interaction dynamics at this finer temporal scale, at which social ties are actually maintained, is a key to the better understanding of the evolution of egocentric networks and the emergent structural correlations in the social network (as we will discuss in Chapter 3). As a demonstration, Fig. 2.10a illustrates mobile phone call patterns in an egocentric network, where the overall activity of the ego (green row) and activities on separated edges with three friends (orange rows) are presented. By looking at this picture we can draw three important observations: (a) the communication dynamics of the central ego (green) is evidently bursty with heterogeneous inter-event times and broad distribution (see Fig. 2.10d) with exponent $\alpha = 0.7$ (for SMS see [203]); (b) The amount of communication efforts is not evenly distributed among ties (orange), but some ties carry the wast majority of interactions, while others are maintained by only a few events. It assigns differences in terms of social tie strengths, arguably associated to various level of intimacy as suggested by Dunbar [45], and leads to heterogeneous link weight distributions on the network level; Finally, (c) correlated events form trains in bursty periods. The distribution of trains in the egocentric network can be explained by two competing hypothesises. On one hand, correlated event trains of the ego may evolve on single links (as seen on the zoom shown on panel Fig. 2.10b), which suggests that bursty periods are actually induced by dyadic interactions. On the other hand, as demonstrated in Fig. 2.10c, bursty communication periods of an ego may involve multiple peers, suggesting that bursty patterns are potentially induced by collective behaviour, the effort of a group e.g. to organise an event or to process information. As next [203], our aim will be to empirically decide between these hypothesises by analysing DS1 mobile phone call and SMS datasets with the previously introduced characteristic methods of bursty phenomena.

More precisely, we are going to use the $P(E)$ bursty train size distribution to indicate how trains are distributed in the egocentric topology. At first, let's concentrate on ego-initiated events of outgoing voice calls and SMS'. In case we consider entire event sequences of egos, which combines all of their communications on any links, the $P(E)$ distributions (shown for calls in Fig. 2.10b with solid lines) appear approximately as a power-law function with exponent $\beta = 4.2$ (for SMS not shown [203]). This behaviour is remarkably different from the scaling of the corresponding reference distributions, where inter-event times were randomly shuffled over the whole data (see exponential distributions with solid lines in Fig. 2.10b for calls). Based on this node centric view, intuitively one would assume that correlated outgoing communications of an individual may serve the information processing or organisation of a group [247, 118], resulting events grouped in trains

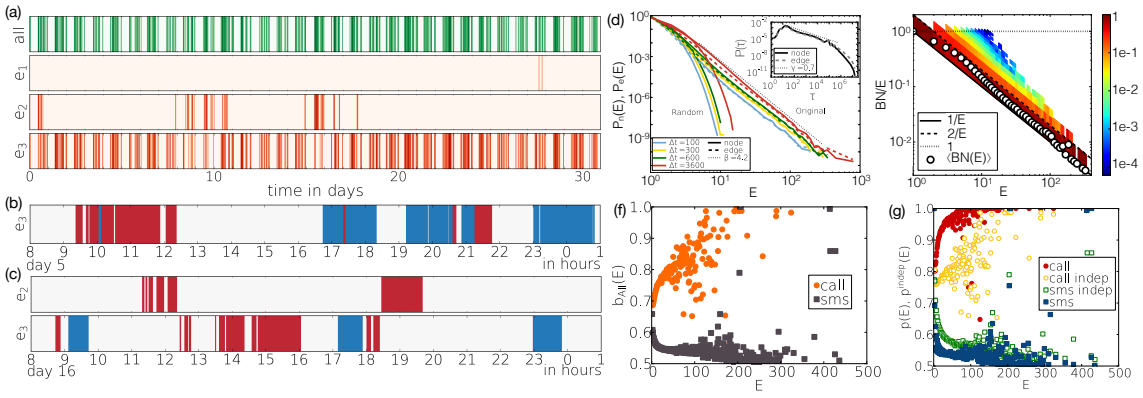


Figure 2.10: (a) Overall activity of the ego (green) and its neighbours (e_1, e_2, e_3). Darker colour scales with call number per hour. (b,c) In- (red) and outgoing (blue) calls with length of the corresponding calls. (d) $P(E)$ distributions of outgoing calls of nodes towards all the neighbours (solid lines) and to single neighbours (dashed lines) for various Δt (in second) using the original and inter-event time shuffled (random) sequences. Inset: inter-event time distributions between outgoing events of a user towards all neighbours (solid line) and on a single link (dashed line). (e) The distribution and average of the ratio BN/E for each E train size. Pointed, solid and dashed lines assign limiting cases of 1, $1/$E, and $2/E$. (f) Average b_e edge balance calculated for trains with the same size for calls (orange circles) and SMS (brown squares). (g) Average $p(E)$ balance values for trains of the same size for calls (red circles) and SMS (blue squares). Corresponding independent event trains are shown with yellow circles for calls and green squares for SMS. This figure was published in [203].

directed towards several neighbours. Assuming this mechanism to be dominant, burstiness would appear as the property of a single node or a group of individuals.

Surprisingly, the generic picture seems to be very different. If we assume that the correlated events in trains are directed toward several neighbours, decoupling event trains on single edges should induce an entirely different, less correlated statistics of bursts. However, this is not the case. The $P(E)$ distributions, detected on single edges (shown for calls in Fig. 2.10b with dashed lines), are scaling very similarly and can be characterised by the same exponents as in the node centric case. This suggests that trains of events usually evolve on single links rather among a larger group of individuals. This picture is also supported by the statistics of temporal motifs [223], where motifs involving two individuals are by far the most common ones. The same conclusion can be drawn by counting the number of neighbours BN, whom an individual called in a bursty train of E events. If a user communicates with only one neighbour during a period then the ratio $BN/E = 1/E$, or if each call are directed toward different neighbours than $BN/E = 1$. The distributions of the BN/E ratios for each E train size (shown in Fig. 2.10e for calls, for SMS not shown [203]) indicates that on average only one or two people are called in a bursty train independent of its size.

Next let's have a look at the direction of event trains. Are they balanced or contain events dominantly initiated by one partner? Do voice calls and SMS' are different from this point of view? First, let's calculate the overall communication balance over the entire observation period for each edge e as $b_e = max(N_A, N_B)/(N_A + N_B)$, where N_A (N_B) are the total number of calls from A to B (B to A). Hence b_e can vary between $1/2$ (completely balanced) and 1 (completely imbalanced, dominated by one of the participants). We use this measure to compute the weighted average *communication balance of trains on single edges* with size E evolving as:

$$b_{All}(E) = \langle b_e \rangle_E = \frac{\sum_e n_e(E) b_e}{\sum_e n_e(E)} \tag{2.15}$$

where $n_e(E)$ is the number of trains of length E on edge e. For SMS (brown squares in Fig.2.10.f) b_{All} is converging to $1/2$ for larger E sizes thus longer trains are more and more balanced in this

case. This can be explained by the uni-directed information flow in case of SMS forcing mutual discussions to be more balanced, as it has been also argued by Wu et al. [380]. However, this constrain does not apply for the mobile calls (orange points in Fig.2.10.f) where information can flow in both directions during a call. Here b_{All} reflects strongly unbalanced communication as it increases towards 1 for trains with larger E.

However, the $b_{All}(E)$ average overall balance does not reflect evidently the communication balance evolving in single trains. Hence we define the *communication balance within a train* on an edge e as $p_e(E_m) = max(n_A, n_B)/(n_A + n_B)$. Here $p_e(E_m)$ is the balance of the $m-$th train of length E_m on edge e connecting A and B, and n_A (n_B) denotes the number of events initiated by A (resp. B) towards B (resp. A) in that train; $E_m = n_A + n_B$. Averaging over trains of the same size $p(E) = \langle p_e(E_m)\rangle_E$ gives an estimate for the average communication balance in trains of size E (note that in this case different $p_e(E_m)$ values can evolve even for trains on the same edge e). This has to be compared to the reference case, where trains of a given size follow the overall balance b_e of the actual link. This can be calculated as

$$p_e^{indep}(E_m = E) = \frac{1}{E}\sum_{i=0}^{E}\left(\left|\frac{E}{2} - i\right| + \frac{E}{2}\right) b_e^i (1-b_e)^{E-i}\binom{E}{i}, \tag{2.16}$$

where the first term after the summation weights the binomial distribution taking into account that the imbalance can evolve in both directions, i.e., parallel or antiparallel to the imbalance of the edge. As $p_e^{indep}(E_m = E)$ depends only on b_e and E, the average can be taken as $p^{indep}(E) = \langle p_e^{indep}(E_m)\rangle_{E\ trains}$ to get a reference estimate of and independent process.

Fig. 2.10.g shows $p(E)$ and $p^{indep}(E)$ for both voice calls and SMS messages. Interestingly, large difference is observed between $p(E)$ and the corresponding $p^{indep}(E)$ measures. It suggests that call trains (red points) are more unbalanced than one would expect from the overall communication balance of the link, caputured by the independent processes (yellow circles). At the same time for SMS the contrary is true as trains (blue squares) are much more balanced than one would derive from independent processes (green squares) reflecting the b_e balance of the corresponding edges. This demonstrates real correlations between events of the same train and suggests different correlated mechanisms behind call and SMS dynamics.

2.5 Models of bursty human phenomena

As we have already explained in the introduction of this Chapter (Section 2.1), there are three main modelling frameworks, which have been proposed to explain the origin of bursty patterns in human dynamics. One framework provides a variety of models using *priority queues* [39, 287, 361]; a second one is based on the assumption that consecutive actions of individuals are independent and can be modelled by *Poisson processes* with alternating time scales [255, 254]; while the third direction assumes strong local correlations modelled by *memory processes* [360, 160], *self-excited point processes* [259, 184], or *reinforcement mechanisms* [203, 365], etc. In the following, we are going to discuss two models [203, 202], proposed by me and colleagues, which belongs to the third modelling framework and employs reinforcement processes to model simultaneously temporal heterogeneities, bursty trains, or communication balance. While these models aim to explain phenomena observed on the individual or dyadic level, later in Section 3.3 we will discuss other models [195, 218], which address the emergence and effects of bursty phenomena on the collective level.

2.5.1 Model of individual bursty dynamics with event trains

According to the Decision Field Theory of psychology [71], each decision to perform a task can be interpreted as a threshold phenomenon, as the stimulus of the task has to reach a given threshold

level to be chosen from the enormously large number of possible actions. This theory lays behind one possible interpretation of bursty behaviour, where the dynamics of an individual performing events of one certain task, like writing emails or printing, goes through active and inactive periods. An active state is initiated once the importance of the actual task overreach a certain level, which after the person performs a bursty cascade of events, while otherwise doing something else which in turn appears as long inactive periods in the actual observation. However, in active periods events are not independent from each other but form long bursty trains as we have already demonstrated in Section 2.2.1. Such dynamics can be explained by memory driven processes and modelled by reinforcement mechanisms as we explain next [202].

Memory process

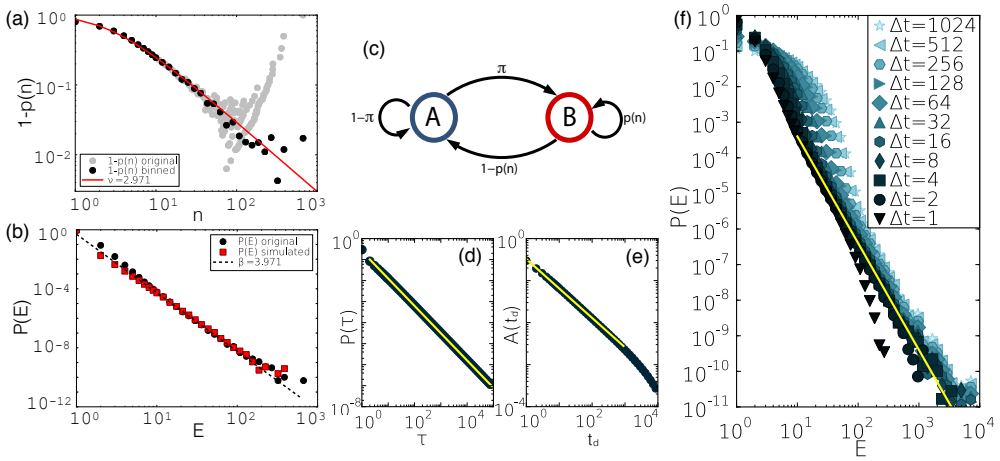


Figure 2.11: (a) The $1-p(n)$ complement of the memory function measured from the mobile call sequence with $\Delta t = 600$ second and fitted with the analytical curve defined in Eq. 2.17 with $\nu = 2.971$. Grey symbols are the original points, while black symbols denotes the same function after logarithmic binning. (b) The $P(E)$ distributions measured in real and in modelled event sequences together with a power-law function with exponent derived from Eq. 2.17. (c) Transition probabilities of the reinforcement model with memory. (d) Logarithmic binned $P(\tau)$ inter-event time distribution of the simulated process with a fitted exponent $\gamma = 1.3$. (e) The average logarithmic binned autocorrelation function with a maximum lag $t_d^{max} = 10^4$ and fitted exponent $\alpha = 0.7$. (f) Logarithmic binned $P(E)$ distributions of the synthetic sequence with window sizes $\Delta t = 1...1024$ and fitted exponent $\beta = 3.0$. Simulation were averaged over 1000 independent realisations with parameters $\mu_A = 0.3$, $\mu_B = 5.0$, $\nu = 2.0$, $\pi = 0.1$ and $T = 10^9$. For the calculation we chose the maximum inter-event time $\tau^{max} = 10^5$. This figure was published in [202].

The correlations taking place between consecutive bursty events can be interpreted as a memory process, allowing us to calculate the $p(n)$ *probability that the individual will perform one more event within a* Δt *time frame after it executed n events previously* in the actual train. This probability can be written as $p(n) = \frac{\sum_{E=n+1}^{\infty} P(E)}{\sum_{E=n}^{\infty} P(E)}$, thus its functional form is entirely coded in the train size distribution. If we assume that the actual train size distribution scales as $P(E) \sim E^{-\beta}$ (as already discussed in Eq.2.11) the memory function appears as

$$p(n) = \left(\frac{n}{n+1}\right)^{\nu} \qquad \text{where} \qquad \beta = \nu + 1 \tag{2.17}$$

scaling relation is expect to hold as derived in [202]. To demonstrate that Eq. 2.17 holds for real systems, we measured the $P(E)$ distribution for mobile calls in $DS1$ with $\Delta t = 600$ seconds and derived the corresponding $p(n)$ function. In Fig.2.11.a we show the $1-p(n)$ complement of the empirical memory function, with strong finite size effects (grey dots) and the same function

after logarithmic binning (black dots) on which we fit the theoretical memory function defined in Eq 2.17 using a non-linear least-squares method with only one free parameter, ν. As seen in Fig.2.11.a, we find that the best fit (red line) offers an excellent agreement with the empirical data with $\nu = 2.971 \pm 0.072$. Using the above mentioned exponent relation, this way we can estimate $\beta \simeq 3.971$, which is close to the empirical value $\beta \simeq 4.1$ already reported in Table 2.1 and Fig 2.5a. To validate this approximation we generated bursty trains of 10^8 events by using the theoretical memory function $p(n)$ (defined in Eq. 2.17) with exponent $\nu = 2.971$ and compared the scaling of the generated $P(E)$ distribution to the corresponding empirical result. Results in Fig.2.11.b evidently demonstrate a good match between the simulated and empirical results, thus in turn they validate the chosen analytical form for the memory function (for more results see [202]).

Reinforcement model of bursty dynamics

Based on the above observations, we assume that the activity of an individual performing a task can be described with a two-state model, where a person can be in a normal state A, executing independent events with longer inter-event times, or in an excited (bursty) state B, performing correlated events with higher frequency. In this model we assume that inter-event times are determined by a reinforcement process, which dictates that the longer the system waits after an event, the larger the probability that it will keep waiting [331, 389]. Thus our two-state model is strongly non-Markovian as the timing of its events depends on the current and past states of the system. More precisely, given that our model system performed its last event τ time ago, the probability that it will wait one time unit longer without performing the next event is given as

$$f_{A,B}(\tau) = \left(\frac{\tau}{\tau+1}\right)^{\mu_{A,B}} \tag{2.18}$$

where μ_A and μ_B control the reinforcement dynamics and the characteristic inter-event times in state A and B, respectively.

Finally, the model is defined as follows (for schematic demonstration see Fig.2.11.c, while an algorithmic description see Alg. 1): first the system performs an event in a randomly chosen initial state (line 2 in Alg. 1). If the last event was in the normal state A, it waits for a time induced by $f_A(\tau)$ (line 7), after which it switches to state B with probability π and performs an excited event (line 9 and 10); or with probability $1-\pi$ stays in the normal state A and executes a new normal event. In the excited state the inter-event time for the actual event comes from $f_B(\tau)$ after which the system executes one more excited event with a probability $p(n)$ (see Eq. 2.17 and line 14 in Alg. 1) that depends on the n number of excited events since the last normal event; otherwise it switches back to a normal state with probability $1-p(n)$ (line 16).

The results of the simulated model process, summarised in Fig. 2.11 and Table 2.1, indicates that the emergent inter-event time distribution appears with strong inhomogeneities (see Fig.2.11.d). It can be approximated by a scale-free function with exponent $\alpha = 1.3$, which satisfies the expected exponent relation $\alpha = \mu_A + 1$, similar to the one derived in Eq. 2.17. Beyond temporal heterogeneities we detect emergent long temporal correlations reflected by the power-law tail of the autocorrelation function (see Fig.2.11.e). It can be characterised by an exponent $\alpha = 0.7$, which also satisfies the relation $\alpha + \gamma = 2$ (as discussed in Section 2.2 and [202, 357]). Finally, the $P(E)$ distribution appears with a fat-tail for each investigated window size ranging from $\Delta t = 1$ to 2^{10} (see Fig.2.11.f), which can be characterised with an exponent $\beta = 3.0$ in agreement with the expected relation in (Eq. 2.17). Note that the weak Δt dependency of the simulated $P(E)$ can be explained by the merge of correlated long bursty trains and uncorrelated single events which is more common for larger Δt. Finally, once we fix the value of α and β, the emergent γ exponent satisfies the inequality $\gamma < \alpha < \beta$, observed in empirical data (see Table 2.1).

Algorithm 1 Algorithmic description of the reinforcement model of bursty activity trains.

function BURSTY REINFORCEMENT MODEL
  $\sigma \leftarrow rand(A,B)$
  $n = 1$
  $time = 0$
  while $time < T$ **do**
    if $\sigma == A$ **then**
      $time+ = f_A(\tau)$
      if $rand() < \pi$ **then**
        $\sigma \leftarrow B$
        $n \leftarrow 1$
      $Out : (time, \sigma)$
    else if $\sigma == B$ **then**
      $time+ = f_B(\tau)$
      if $rand() < p(n)$ **then**
        $n \leftarrow n+1$
      else
        $\sigma \leftarrow A$
      $Out : (time, \sigma)$

2.5.2 Model of communication balance of dyadic event trains

In an extended model definition, introduced in [203], we further used the $p(n)$ memory function (in Eq. 2.17) to model the observed cases of communication balance in call and SMS sequences as reported in Section 2.4.2. To introduce this new model let us first concentrate on voice calls. One correlation we observed in Fig. 2.10.f was that longer event trains tend to be more unbalanced, meaning that they are more dominated by events initiated by one of the callers. Keeping in mind that mobile calls enable bidirectional information change, we assume that the observed unbalanced communication in interaction trains reflects the difference in motivation between the communicating partners. If there is a task to solve, which is more important for one party, it gives motivation for him/her to repeate calls until the issue gets settled. This mechanisms can be incorporated into the reinforcement process demonstrated in Fig.2.12.a, and can be summarised in the following way (note that its algorithmic solution is somewhat similar to Alg. 1 thus we do not present it here): We simulate bursty trains, which evolve on a link between a pair of individuals A and B. To initiate a train with a probability equal to b_e we randomly select A or B who then perform one event towards the other agent and we set the actual train size to $n = 1$. The decision about the next event is carried out in two steps. First we decide with the probability in Eq.2.17 whether to perform one more event in the train or initiate a new train otherwise. If the train should be continued the probabilities that the next event is initiated by A or B are

$$p_\sigma(n|\sigma_1) = \frac{n}{n+1} \quad \text{or} \quad p_\sigma(n|\neg\sigma_1) = 1 - \frac{n}{n+1} \tag{2.19}$$

where $\sigma \in \{A,B\}$. Here $p_\sigma(n|\sigma_1)$ denotes the probability that the nth event of the actual train is performed by the same user who initiated the train at $n = 1$, while $p_\sigma(n|\neg\sigma_1)$ is the probability that the other agent initiates the event. Consequently, the longer a train evolves, the larger is the probability that the agent, who initiated the actual train, will initiate the next event. Eq.2.17 and Eq.2.19 may capture the entangled mechanisms of reinforced motivation of an individual, which is induced by the effort already invested in the actual series of calls to successfully solve a task with the other partner.

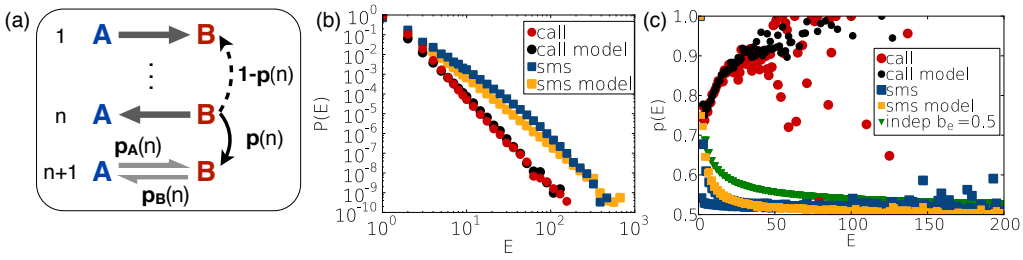


Figure 2.12: (a) Illustrative definition of the communication balance model, simulating events between two nodes A and B. The dynamics and direction of the events are determined by probabilities $p(n)$, $p_A(n)$ and $p_B(n)$ defined in the text. (b) $P(E)$ distributions of empirical call trains (red circles) with $\Delta t = 600s$ on edges with $0.5 \leq b_e < 0.55$ and in corresponding model trains (black circles). The same functions are shown for SMS trains (blue and yellow squares accordingly). (c) Balance values calculated for empirical call (red circles) and SMS (blue squares) trains and in corresponding model processes (red and yellow symbols). Balance values of independent trains are also shown (green triangles) calculated by using Eq.2.16 with $b_e = 1/2$. This figure was published in [203].

In case of SMS sequences the mechanism for developing strong balance in bursty trains is different. There, in single events information can pass only one way and consecutive events in a train usually have reversed direction, possibly forming strongly balanced conversations. To simulate this behaviour we use the above defined generative process but we select the direction of the actual event differently. Here we assume that the direction of an event conditioned only on the orientation of the previous event, and can be determined by the conditional probabilities

$$p_\sigma(n|\neg\sigma_{n-1}) = \frac{n}{n+1}, \quad p_\sigma(n|\sigma_{n-1}) = 1 - \frac{n}{n+1} \tag{2.20}$$

where $\sigma \in \{A,B\}$ and $p_\sigma(n|\neg\sigma_{n-1})$ denotes the probability to choose the opposite direction for the nth event compared to the one in the $(n-1)$th step. Accordingly $p_\sigma(n|\sigma_{n-1})$ denotes the probability of choosing the same direction as the previous event. This way, the longer a train evolves, the larger is the probability to revert the direction of consecutive events and consequently to generate more balanced train.

The emergence of enhanced balance/imbalance in trains can be evidently checked on links where the overall communication is completely balanced and the communication balance of trains comes only from actual behavioural differences. To do so we set $b_e = 1/2$ and compare the modelled results to averages calculated for similar empirical trains. To analyse real trains we select edges from the mobile call network with overall balance $0.5 \leq b_e < 0.55$ (there are $115,277,534$ calls and $69,288,504$ SMS on such links) and after detecting trains we calculate the corresponding $P(E)$ distribution and $p(E)$ function (defined in Section 2.4.2). As expected and shown in Fig.2.12.b, the size of call trains (red circles) detected with $\Delta t = 600s$ and SMS trains (blue squares) with $\Delta t = 300s$ are distributed broadly with characteristic exponents $\beta = 4.6$ and $\beta = 3.5$, accordingly. Using these exponents as parameters we modelled event sequences to simulate calls and SMS trains with the same number of events and corresponding ν exponents deduced from β according to Eq.2.17. The $P(E)$ size distributions of model call trains (black circles) and model SMS trains (yellow squares) collapse surprisingly well on the corresponding empirical functions, as shown in Fig.2.12.b.

At the same time, in Fig. 2.12.c the $p(E)$ balance functions calculated for the limited event sets on fully balanced links show surprisingly similar behaviour to the overall averages (seen in Fig. 2.10.g). This demonstrates that even an overall balanced link, strong communication imbalances appear due to local correlations within one bursty train. This is even more striking if we compare the empirical P(E) curves to the corresponding independent one (green triangles in

Fig.2.12.c) which was generated using Eq.2.16 with $b_e = 1/2$. Moreover, in Fig. 2.12, the average $p(E)$ balance functions emerging without parameters for model trains are in surprisingly good agreement with the empirical observations. Consequently, the assumed mechanisms defined in Eq. 2.19 and 2.20 are capturing rather accurately the salient features of the dynamics of directed human communication through phone calls and SMS. The only discrepancy is for the $p(E)$ values of short SMS trains, where the empirical data show an even-odd effect, which is not reproduced in the model. This indicates that possibly other mechanism may be present in the communication sequence what are not considered in this modelling study.

2.6 Conclusions

In this Chapter, I summarised my most important results in one of my main research domains on bursty human dynamics. After a brief overview of the field, first I introduced some basic characteristic measures, some of them defined by me and colleagues, which are commonly used to quantify heterogeneous patterns in human dynamics. Subsequently, I systematically walked through a series of studies I published over the last years for the advanced characterisation, observation, and modelling of bursty human dynamics.

Due to the diverse experiences, broad overview, and devoted interest towards this field, together with Dr. Hang-Hyun Jo and Pr. Kimmo Kaski, we recently took the timely opportunity to write a monograph book, entitled as ”Bursty Human Dynamics“, to review all relevant knowledge on the field cumulated over the last decade. This book has been published by Springer in January 2018.

Finally note that some of my studies published on the system-level observations, modelling, and effects of bursty dynamical patterns are not mentioned in this Chapter as they land close to field of temporal networks, which is the topic of the coming Chapter.

3. Temporal Networks

3.1 Introduction

The success of Networks Science is built on the operational representation of complex systems as graphs, which in turn can be quantified, observed, and modelled for the better understanding of emerging phenomena [278]. From the early years of the field a common simplifying assumption has been taken to neglect that real networks may evolve in time, and may consist of nodes and links of different types. Although these approximations were obviously vague in various cases, yet the static/monolayer network approach was extremely successful to obtain crucial knowledge about several empirical systems. However, after a decade the field became advanced enough to overcome these limitations and to develop the contemporary domains of *temporal networks* to consider the dynamic nature of interactions [173, 258], and *multilayer networks* to take into account the multiplicity of interaction types [216, 55]. These recent developments were fuelled by the ever increasing network data available, which more-and-more commonly comes in the form of time-stamped interaction sequences and/or with metadata recording details on nodes and links. Advancements in data collection and data sharing together with the novel theoretical foundation of these domains help us to break the glass ceiling of the static/monolayer network picture and push us to think about networks as dynamical systems with various types of interactions and agents. In this Chapter we concentrate on temporal networks as I mainly contributed to the theoretical, methodological, and empirical foundation of this domain. First we lay down some general thoughts on temporal networks, which after we will introduce a set of representations and characteristic measures which are necessary to understand the body of this Chapter. Subsequently, we will discuss four main directions of my contributions on to the system level characteristic, random reference models, higher order representations and generative modelling of temporal networks.

3.1.1 Static vs. Temporal Networks

Most network structures are the results of some emergent phenomena, consequently they continuously evolve in time, or dynamical evolution was necessary at some point of their existence. This way, by only looking at the static description of their aggregated final or actual state we may obtain limited understanding about their "morphogenesis" and emergent properties. A good example

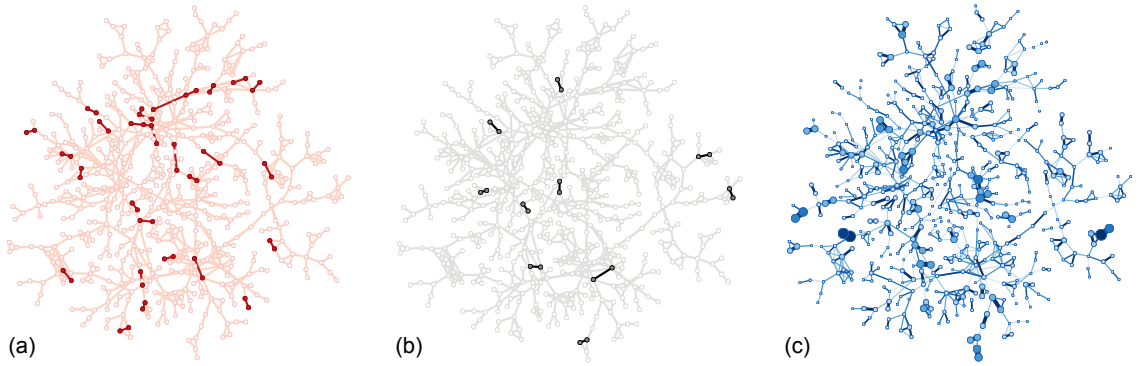


Figure 3.1: Dynamics of a mobile call network. Panels (a), and (b) show calls over 3 hours between people in the same town at two different time stamps. Panel (c) presents the backgrounding weighted social network structure, which was recorded by aggregating interactions evolved between people during 6 months. Node size and colours describe the activity of users, while link width and colour represent weight. This figure was published in [206].

is the commonly observed degree heterogeneity, which has been argued to emerge due to some reinforcement mechanisms (preferential attachment, Matthew process, etc) driving the interaction dynamics in a network [40]. Other examples are degree correlations, weight heterogeneities, network communities, etc., which all can be observed in a static network as emergent properties, due to mechanisms driving the underlying network dynamics at the first place. Temporal networks contributes to their understanding by studying the network dynamics at the level of rapid recurrent interactions between nodes [173]. As we will see later, this level of description can be used to identify important mechanisms and correlations between single events of interactions [195, 233, 206], which leads to emergent heterogeneous properties on the aggregated scale (as demonstrated in Fig. 3.1). Taking interactions as static links between nodes subsequently entails that information can flow between the connected peers at any time [206]. This is evidently not true in many cases, like in human communication networks [195, 217], where information can be passed between nodes only at the time of interactions and only between the interacting nodes. This way, information flow, and as a consequence the emergence of any macroscopic phenomena on networks, are limited by the emerging time respecting paths determined by the the timing, direction, and ordering of events [217]. This puts network analysis in a rather different perspective, as all characteristic network properties, centrality measures, or structural properties become time dependent [388, 226]. On the other hand, better understanding of time respecting paths opens the door to more realistic simulations of dynamical processes, like epidemic or information spreading, diffusion processes, opinion dynamics, or synchronisation, which are actually crucially altered by the time-varying nature of interactions (for further discussion on this matter see Section 3.5.2).

3.1.2 Time-scales of network dynamics

Taking an observer point of view, a network can dynamically evolve in various temporal scales relative to the observation frequency [310]. The typical cases are demonstrated schematically in Fig. 3.2, where we fix the observation frequency and vary the relative time-scales between two extremes. On one end of this scale we find the so called *quenched networks*, which are invariant in time or evolve on such a slow pace that appears to be invariant from the observer point of view. Static networks are belonging in this category with frozen nodes and links, which are always present in the structure. Once we consider that nodes and links may be created and deleted in the network, but yet on a slower temporal scale as the observation, we arrive to the *evolving network* picture [17]. Best examples for this representation are social networks where individuals may born

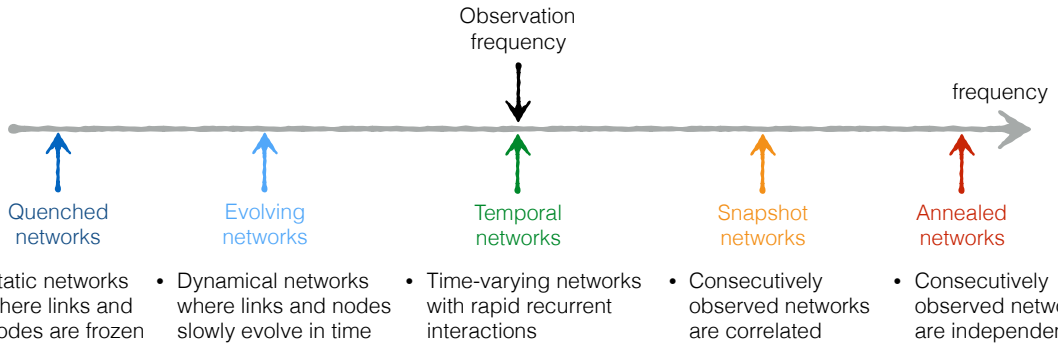


Figure 3.2: Relative times-scales of network dynamics as compared to the time-scale of observations.

and die and social ties are created and broken but all being present for longer periods lasting for several observations. Other examples are the internet, or other infrastructure networks. Once the observation frequency closely matches the temporal scale on which the network is evolving we arrive to *temporal networks* [173]. Here we consider rapid and potentially recurrent time-varying interactions between nodes like emails or phone calls in social communication networks, or sexual contact networks. Note that temporal networks contain all information provided by the other network pictures evolving slower than the observation frequency. One can take first and last observations of interactions between nodes to reconstruct the evolving network representation, while aggregating interactions over time would give us the static network description of a time-varying network [173]. However, we loose this advantage once the network is evolving faster than we can observe. In this case observations provide *snapshot networks*, which aggregate the structure of the time-varying network over short consecutive periods. If the time-scale difference is still not too large, consecutive snapshots are not entirely independent from each other and can be used as an approximate method to represent temporal networks. For a good example see Fig. 3.1.a and b where two snapshots of a mobile phone call network is shown. On the other hand, once the network is evolving on a way faster time-scale than we can observe we arrive to the *annealed network* picture in which case consecutive observations of the network are weakly correlated or entirely independent [284, 125]. Neural networks in the brain are good example for this case, where imaging technologies have not reached yet the temporal (neither spatial) resolution necessary to observe the precise electric signals running between neurons. Also note that the so called *mean-field approximation* becomes valid in this extreme.

3.1.3 Representations of temporal networks

In order to study temporal networks we need to introduce a representation, which can be further used to define general measures and which maps between a mathematical description and the common data collection format of time-varying interactions. The most straightforward description, which we will consistently use later on, is the event list (also called event-based) representation. To introduce, let us consider a temporal network as a set $G_t = (V, E_t, T)$ defined as sets of vertices $v \in V$, and edges defined as a set of events $E_t \subset V \times V \times [0,T]$ over a time period T. This way a temporal network is described as a sequence of events, which in their simplest form appear as triplets (u,v,t) indicating an interaction between nodes u and v at time $t \in [0,T]$. Note that an event can be directed or undirected depending on whether we consider $V \times V$ as an ordered or unordered set. Furthermore, this way of representation allows for both continuous and discrete time description of temporal networks. As data commonly comes in a discrete time format, we chose this as our default assumption and we consider events undirected if not mentioned otherwise. This representation can be further extended by defining events as $E_t \subset V \times V \times [0,T] \times \prod_i A_i^e$ where A_i^e

is a set of meta informations like duration, cost, or the location of participating nodes, which may extend the description of each event depending on the actual data.

Several other representations of temporal networks has been provided over the last couple of years [173, 258]. One represents temporal networks as a discrete time sequence of snapshot networks captured by adjacency matrices $\mathscr{A} = \{A(t)\}$. Each matrix $A(t) = |V| \times |V|$ aggregates all events, which are present between nodes at time t. However, since the ordering of events within one snapshot is not determined, this representation does not contain all information about the original temporal structure, and the information loss depends on the relative time-scales of the snapshot time window and the evolving network. Furthermore, the notion of duration in this representation is not straightforward. If events have shorter duration than the sampling frequency we may loose information again. Or in the contrary, if events have long duration, they appear as multiple events in several consecutive time windows. Another recently proposed representation is so called *link streams* of a temporal network [232]. This description aims to provide the most general representation of dynamic networks by using stream graphs defined as a $S = (V, W, E, T)$, where W being the set of temporal nodes $W \subset V \times [0, T]$. In case we assume that W is time invariant, i.e., all nodes are present always, we receive the definition of link streams which provides a similar description with the even list representation discussed above. As follows we are going to mostly apply the even list representation, while in Section 3.4.2 we will propose a new lossless representation of temporal networks as static weighted directed acyclic graphs.

3.1.4 Some characteristic measures of temporal networks

There are several ways to extend the definition of characteristic measures of static networks (e.g. ones mentioned Section 1.2.1) for temporal structures. Here, without aiming a complete review, we will summarise some concepts and metrics, which will be frequently used in the forthcoming Chapters for the analysis of time-varying networks. For a more complete list of definitions we refer the reader to recent review articles and books like [173, 258, 232].

Dynamical measures of characteristic metrics

The first thing to notice in case of temporal networks is that characterising metrics, like degree, strength, weight, clustering coefficient, or centrality measures may vary in time. However, despite their timely variance, their distributions or average values may reach stationary values. The most straightforward way to capture their dynamics is via assuming a time aggregation process. There, initially at $t = 0$, we have an empty graph with N nodes and take each event in a timely order to build an aggregated static network $G_{[0,t]} = (V, E_{[0,t]})$. At time t this static structure will have links induced by interactions appeared between $[0, t]$ and can have link weights defined as the number of interactions, the sum of event durations etc. Measuring the average or the distribution of characteristic network metrics in the aggregated network $G_{[0,t]}$ as $t \to T$ give us some information about the evolution of the network and the emergence of its stationary properties [226]. Note that in the limit $t = T$ the representations $G_{[0,T]} \equiv G$ are equivalent. There are several other ways to measure general characteristics, like via the analysis of networks observed in consecutive discrete aggregated snapshots. However, these methodologies are not discussed here as they are not necessary for the understanding of the rest of the thesis.

Path-based measures

Some of the most important characters of temporal networks build on the concept of temporal paths (also called time-respecting paths). Paths in temporal networks have rather different properties as compared to static networks. Most importantly, they fully determine the outcome of any collective phenomena as they constrain possible information flow between nodes. However, before we ground the definition of temporal paths we need to introduce some other concepts related to event adjacency and temporal walks. In the following we are going to use the event based representation of an

undirected temporal networks, but all definitions are generalisable for other representations as well.

Event adjacency defines a directed relationships between events which are performed on neighbouring links. More precisely, taking two events $e_1 = (v_i, v_j, t_1)$ and $e_2 = (v_j, v_k, t_2)$ from a temporal network (with $v_i, v_j, v_k \in V$ and $t_1, t_2 \in T$), the event e_1 is adjacent to event e_2 if they share at least one common node and $t_1 < t_2$. Next, let's imagine a walker on the temporal network, which can pass between nodes but only at the time of their interactions while respecting the ordering of events, i.e., it can use only events from the future. Then, a temporal walk between two nodes v_0 and v_n is defined as a sequence of $\{(v_0, v_1, t_1), (v_1, v_2, t_2) \dots (v_{n-1}, v_n, t_n)\}$ of adjacent events, where $t_i < t_{i+1}$ $(0 \leq i \leq n-1)$. In general, a temporal walk depends on the time t_1 when it starts, it can visit a node multiple times, and its length is defined as $t_n - t_1$. Furthermore, a temporal walk is non-symmetric (even in an undirected temporal network), non-transitive, and cannot be infinitely long (for further explanation of these properties see [258]). Subsequently, we can define the concept of *reachability*, i.e., a node v_j is reachable from a node v_i if there exists a temporal walk from v_i to v_j. We call a temporal walk between two nodes to be a *temporal path* if the walker visits one node maximum once (thus no node is visited multiple times during the walk).

Finally, we introduce the concept of *connected components* in temporal networks, which is different from the corresponding static network definition due to the non-symmetric property of temporal paths. A connected component in a temporal network is a maximal set of nodes in which each pair of nodes is temporally connected. Following this definition one can define weakly and strongly connected components, but one needs to keep in mind that a node can participate in multiple connected components depending on the temporal paths in which it participates. This way connected components do not provide a well defined node-partitioning in temporal networks. In terms of computation, the complexity to find all connected components in a temporal network is an $\mathcal{O}(N^2)$ problem compared to the static network solution with $\mathcal{O}(M)$ complexity. Finding all temporal paths in a temporal network has been an *NP*-complete problem. However, in a recent paper [217] we provided some better algorithmic solution what will discuss in Section 3.4.2, where we will introduce a new, higher-order representation of temporal networks together with new definitions of connected components.

3.2 System level characterisation

3.2.1 Aggregation time for social communication networks

After this brief introduction to temporal networks first we will discuss some results characterising time-varying networks on the system level. As we have discussed earlier, temporal aggregation of time-varying interactions provides a static, weighted network representation, which contains no temporal informations, but codes the underlying structure. This aggregated structure may have various heterogeneous properties and may emerge with communities and other consequences of structural correlations. However, when taking a sequence of time-varying interactions it is not clear how long one needs to aggregate in order to capture the most important static network properties.

Next, we address this question by monitoring and analysing the features of network structure emerging from aggregation over different time intervals for an empirical data set of mobile phone communication [226]. We are interested in the effects of the aggregation window on the structural features of human communication networks that are known to display dynamics on multiple overlapping time scales. The data comes in the form of a time-stamped sequence of mobile telephone calls between anonymised customers of a Belgian mobile operator for a period of 6 months. Note that although it has been collected in another country, it is very similar to DS1 in Section 1.3.2 (for more about the data see [226]).

The aggregated network possesses all structural heterogeneities typically characterising social communication networks. It appears with a broad degree, weight and node strength distribution as

reported in [226]. As a first approximation, let's characterise its dynamics by simply measuring the evolution of basic network characteristics. As shown in [226], the number of nodes or links, the average node degree or link weight all increase monotonously in time without indicating any characteristic time scale. Moreover, the interaction dynamics clearly displays the usual daily and weekly pattern, and it appears bursty, indicated by the long tail of the distribution of inter-even times measured between consecutive calls on individual edges.

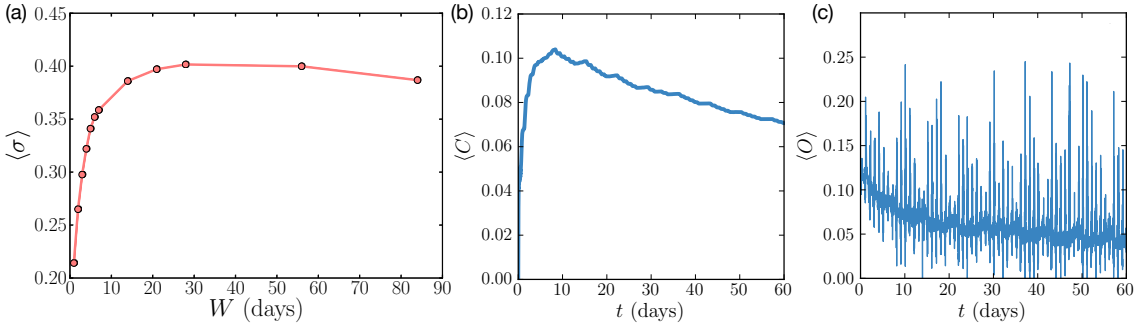


Figure 3.3: (a) The average fraction of links f common to consecutive aggregation windows of duration W. (b) Global clustering coefficient and (c) average final overlap of added edges as a function of aggregation time, for the first 2 months. This figure was prepared by G. Krings and published in [226].

Further observations suggest that links in a social network appear with various time span. While some of them are stable and remain active for prolonged periods of time, others exist or can be detected only within limited time periods. In reflection, the aggregation window length should obtain a representative, "backbone" networks that capture the stablest connections in the system? Thus, we compare the similarity of networks aggregated for different periods of time when the observation period is divided into multiple consecutive aggregation windows. We then calculate the similarity σ of two consecutive networks $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$ as

$$\sigma(G_1, G_2) = \frac{|E_1 \cap E_2|}{|E_1 \cup E_2|}, \tag{3.1}$$

such that $\sigma = 1$ if the networks are the same, and $\sigma = 0$ if they share no links. In Fig. 3.3a, the average similarity σ, computed for different durations W, indicates small similarity if the windows are very short, as the networks are very sparse with only a few common links. Then, the similarity increases with increasing window duration, reaching a maximum at ~ 30 days; subsequently, the similarity begins to decrease slowly as the aggregation process captures more and more of spuriously appearing very weak ties or random links.

To better understand the network evolution it is important to learn about the characteristics of links that emerge early on in the process. As we have already discussed in Section 1.2.2, the Granovetter hypothesis suggests that link weights correlate with the network topology such that high-weight links are associated with dense network neighbourhoods, whereas low-weight links connect such neighbourhoods. This is directly related to the presence of community structure [126] in social networks; links within communities are stronger and have higher-than-average weights [339]. For the network aggregation, this means that clusters and communities are likely to appear early on in the process. In order to investigate this effect, we measure the evolution of the global clustering coefficient $C(t)$ and the average final overlap values $\langle O \rangle(t)$ of added links (both defined in Section 1.2.1) as the function of aggregation time. As seen in Fig. 3.3b, the clustering coefficient does indeed increase initially rapidly, and then decreases after a peak at around $t = 9$ days. This decrease can be attributed to the weak links observed later in the process: those links contribute less frequently to triangles. For the case of the average overlap in Fig. 3.3c, the early observed links

have on average a higher overlap than those added later; the final overlap is a decreasing function. Hence, even when the aggregation times are short, the networks capture features of the community structure of the final aggregated networks. Interestingly, the overlap also shows a strong circadian and weekly pattern – its highest peaks correspond to the early morning when the overall call rate is very low. Thus, if calls are made during these hours, they are likely to be targeted towards people in the strongest clusters of friends and family.

In order to illustrate the network growth, we have visualised small subnetworks corresponding to different aggregation times t of people from a single town (see Fig. 3.4). Panels a) to d) of Fig. 3.4 show that at the early times of the observation, those edges appear mostly which participate in triangles in the final aggregated network (coloured in red). These edges are the ones forming communities and clusters. On the other hand, not all intra-community edges are discovered early on; rather, those links that appear early are associated with communities with a high probability.

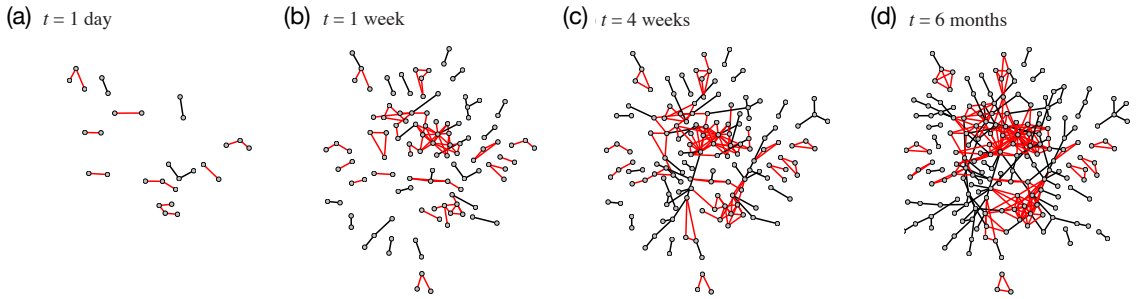


Figure 3.4: Series of aggregated networks with a growing aggregation interval. Networks here represent small subnetworks of individuals from a single postal code. Links that participate in triangles in the final 6-month aggregated network are coloured red, while the rest are black. This figure was prepared by G. Krings and published in [226].

Further investigating the statistical distributions of node degrees and link weights, we observed (not shown here but presented in [226]) that after a short initial transition period, sufficient amount of data has been collected to correctly estimate the stationary degree distribution of the network, and the shape of the distribution remains similar for longer intervals longer than 3 weeks. The weight distribution displays a slightly slower convergence, and is still slowly changing even for aggregation intervals of around 3 weeks.

3.2.2 Entropy of Dynamical Social Networks

To characterise temporal networks on the system level we turn to a more complex measure capturing the entropy of time-varying information coded in the interaction dynamics of individuals. Network entropy has been earlier introduced for static networks [116, 52, 23], but its first definition for temporal networks was proposed by us in [388]. We have seen through several earlier examples that human activities are commonly bursty and not Poissonian and modulated by periodic daily (circadian) or weakly patterns. To understand the convoluted effects of these characters, our question here is: How much can humans intentionally change the statistics of social interactions and the level of information encoded in the dynamics of their social networks? To answer this question, through the analysis of a mobile phone-call network (DS1 in Section 1.3.2), we show that the entropy of dynamical networks is able to quantify the information encoded in the dynamics of time-varying interactions.

The entropy measure

Here we introduce the entropy of dynamical social networks as a measure of information encoded in their dynamics [388]. We assume to have a quenched network G formed by N agents and we allow

a dynamics of interactions on this network. If two agents i, j are linked in the network they interact at each given time giving rise to the time-varying network. In the network G, agents $i_1, i_2, \ldots i_n$ can interact in a group of size n. Therefore at any given time the static network G will be partitioned in connected components or groups of interacting agents as shown in Fig 3.5a or b. In order to indicate that an interaction is occurring at time t in the group of agents $i_1, i_2, \ldots, i_n$ and that these agents are not interacting with other agents, we write $g_{i_1,i_2,\ldots,i_n}(t) = 1$ otherwise we put $g_{i_1,i_2,\ldots,i_n}(t) = 0$. Therefore each agent is interacting with one group of size $n > 1$ or non interacting (interacting with a group of size $n = 1$). Consequently at any given time, the condition

$$\sum_{\mathscr{G}=(i,i_2,\ldots,i_n)|i\in\mathscr{G}} g_{i,i_2,\ldots,i_n}(t) = 1. \tag{3.2}$$

has to be valid, where we indicate with $\mathscr{G}$ an arbitrary connected subgraph of G. The history $\mathscr{S}_t$ of the dynamical social network is given by $\mathscr{S}_t = \{g_{i_1,i_2,\ldots,i_n}(t')\,\forall t' < t\}$. If we indicated by $p(g_{i_1,i_2,\ldots,i_n}(t) = 1|\mathscr{S}_t)$ the probability that $g_{i_1,i_2,\ldots,i_n}(t) = 1$ given the story $\mathscr{S}_t$, the likelihood that at time t the dynamical networks has a group configuration $g_{i_1,i_2,\ldots,i_n}(t)$ is given by

$$\mathscr{L} = \prod_{\mathscr{G}} p(g_{i_1,i_2,\ldots,i_n}(t) = 1|\mathscr{S}_t)^{g_{i_1,i_2,\ldots,i_n}(t)} \tag{3.3}$$

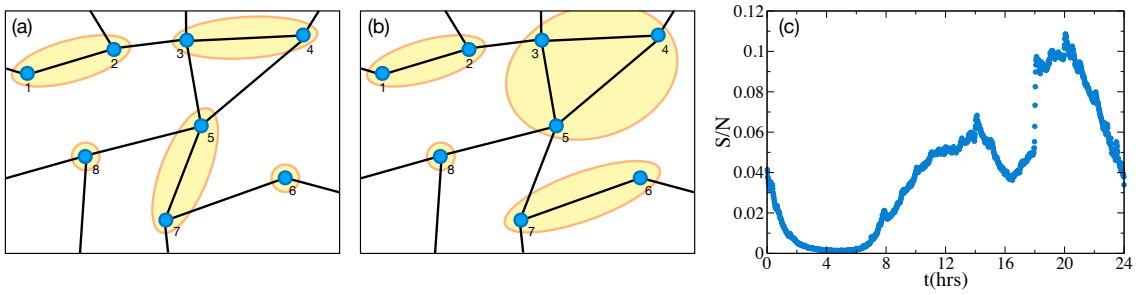


Figure 3.5: The dynamical social networks are composed by different dynamically changing groups of interacting agents. In panel (a) we allow only for groups of size one or two as it typically happens in mobile phone communication. In panel (b) we allow for groups of any size as in face-to-face interactions. (c) Mean-field evaluation of the entropy of the dynamical social networks of phone calls communication in a typical week-day. In the nights the social dynamical network is more predictable. This figure was published in [388].

The entropy S characterises the logarithm of the typical number of different group configurations that can be expected in the dynamical network model at time t and is given by $S = -\langle \log \mathscr{L} \rangle_{|\mathscr{S}_t}$ that we can explicitly express as

$$S = -\sum_{\mathscr{G}} p(g_{i_1,i_2,\ldots,i_n}(t) = 1|\mathscr{S}_t) \log p(g_{i_1,i_2,\ldots,i_n}(t) = 1|\mathscr{S}_t). \tag{3.4}$$

According to the information theory results [93], if the entropy is vanishing, i.e. $S = 0$ the network dynamics is regular and perfectly predictable, while if the entropy is larger, the number of future possible configurations is growing and the system is less predictable. In general, we have to allow the possible formation of groups of any size. However, if we model the mobile phone communication, we need to allow only for pairwise interactions. Therefore, if we define the adjacency matrix of the network G as the matrix a_{ij}, the log likelihood takes the very simple expression given by

$$\mathscr{L} = \prod_i p(g_i(t) = 1|\mathscr{S}_t)^{g_i(t)} \prod_{ij|a_{ij}=1} p(g_{ij}(t) = 1|\mathscr{S}_t)^{g_{ij}(t)} \quad \text{with} \quad g_i(t) + \sum_j a_{ij} g_{ij}(t) = 1, \tag{3.5}$$

for every time t. The entropy is then given by

$$S = -\sum_i p(g_i(t)=1|\mathscr{S}_t)\log p(g_i(t)=1|\mathscr{S}_t) - \sum_{ij} a_{ij} p(g_{ij}(t)=1|\mathscr{S}_t)\log p(g_{ij}(t)=1|\mathscr{S}_t).$$

Social dynamics and entropy of phone call interactions

For demonstration we have analysed the mobile phone-call sequences recorded in DS1. For the entropy calculation we selected 562,337 users who executed at least one call per a day during a week period and we have studied how the entropy of this dynamical network is affected by circadian rhythms. We assigned to each agent $i = 1,2$ a number $n_i = 1,2$ indicating the size of the group where she belongs. If an agent i had coordination number $n_i = 1$ she is isolated, and if $n_i = 2$ she was interacting with a group of $n = 2$ agents. We also assigned to each agent i the variable t_i indicating the last time at which the coordination number n_i has changed. If we neglect the feature of the nodes, the most simple transition probabilities that includes for some memory effects present in the data, is given by a probability $p_n = p_n(\tau,t)$ for an agent in state n at time t to change her state given that she has been in her current state for a duration $\tau = t - t_i$.

We have estimated the probability $p_n(\tau,t)$ in a typical week-day. Using the data on the probabilities $p_n(\tau,t)$ we have calculated the entropy, estimated by a mean-field evaluation (for more on this see [388]) of the dynamical network as a function of time in a typical week-day. The entropy of the dynamical social network is reported in Fig. 3.5c. It significantly changes during the day describing the fact that the predictability of the phone-call networks change as a function of time. In fact, as if the entropy of the dynamical network is smaller and the network is in a more predictable state.

3.3 Random Reference Models

Human actions and interactions are driven by various intrinsic decision mechanisms and influenced by several environmental factors. As the consequence of these convoluted processes various correlations appear at the phenomenological level, which in turn characterise the structure and dynamics of the emerging temporal network. Differentiation of these correlations is not only important to identify and understand the mechanisms underlying the network evolution, but because they have important consequences on the ongoing dynamics processes like in case of spreading phenomena. Spreading processes are relevant for a number of fields and applications ranging from epidemiology of biological viruses to the dynamics of social processes, such as opinion dynamics and information transmission [43]. While certain static characteristics of complex networks work to enhance spreading, such as the small-world or the scale-free properties, it has been shown that the temporal characteristics of links may slow spreading down [362, 195, 217, 268]. These results indicate that dynamical processes cannot necessarily take advantage of topologically characters [292] but they are determined by the combination of structural and temporal characters of the underpinning temporal network.

For static networks, a common way to assess the significance of chosen topological features is to compare their abundance or characteristics against some reference model where the network is randomised. This approach has also been applied in assessing the importance of such features for dynamical processes. The most widely applied reference model is the *configuration network model* [270], where the links of the original network are rewired pairwise randomly. This reference model preserves the original degree sequence but yields networks that are as random as the degree sequence allows. Then, one can assess the significance of topological characteristics of the empirical graph, e.g. by measuring the extent to which the dynamics of some processes differ when they take place on the original networks or the reference ensemble.

Our aim in this Section is to introduce *random reference models* (RRMs) for temporal networks. In this case, the original event sequences are randomised or randomly reshuffled to remove time-

domain structure and correlations [195, 218, 268, 169, 134]. Thus a reference model for an empirical event sequence is a maximally random (micro-canonical) ensemble of event sequences, for which some predefined set of properties are the same as for the empirical sequence. There are various kinds of temporal correlations such like burstiness, causal events, or circadian fluctuations, etc., and thus no single, general-purpose reference model (a 'temporal configuration model') can be designed. Rather, by applying appropriate reference models, one may switch off correlations of selected types in order to understand their contribution to the observed properties. Here, we will briefly introduce the methodological concept, a canonical naming convention and a hierarchical organisation of RRMs [134] and will apply some of these models on an empirical temporal network to demonstrate their potential in identifying important structural and temporal correlations controlling spreading dynamics [195, 218].

3.3.1 Naming and and organisation of RRMs

One can think about random reference models as constrained shuffling methods of interaction sequences, where some property of the original temporal network is retained, otherwise the shuffled sequence is maximally random. Equivalently, such methods sample uniformly from the micro-canonical ensemble of all networks, which have a given set of features constrained to the same value as that of the original data, but have maximum entropy otherwise [218][1]. Consequently, a RRM is determined by the temporal network and a set of selected constrains, which can be related to temporal, structural, or environmental properties, or can be specific to the actual representation of the network. The sequence $\mathbf{t}$ of interaction times, the sequence of $\mathbf{k}$ node degrees, or the $p(\mathbf{w})$ distribution of number of interactions on single links are all good examples for such constraints, which in turn can be used for naming RRMs. For example, the simple random reference model, $P[E]$, constrains only the total number of events E in the network, and permutes all the instantaneous events at random otherwise. On the other hand RRMs can be defined by constraining on multiple characters, or via the intersection of multiple ways of shuffling. The $P[\mathbf{w},\mathbf{t}]$ 'time-shuffling' RRM, randomly permutes the timestamps t of all events, while keeping the participating nodes of the events fixed. Completely equivalently, we may define the timestamp shuffling by constraining the timestamps t and permuting the pairs of interacting nodes among events. Due to the indistinguishability of networks obtained through permutation of event indices or times, both are equivalent to conserving number of interactions on each link $\mathbf{w}$ and the global sequence of interaction times $\mathbf{t}$. This framework enables us to build a taxonomy of existing RRMs, which lists their effects on important temporal network features and (partially) orders them by the amount of features they constrain. This hierarchy allows to apply RRMs so that the fixed features of the original data are systematically reduced. Without going into details, we refer the interested reader to our recent article [134], where we categorically list all identified constrains, and formally introduce the naming convention and a hierarchical organisation of simple and combined RRMs.

3.3.2 Modelling concept and demonstration of RRMs

As it has been discussed in Section 1.1.2, random reference models are school examples of data-driven modelling, where real-world data and synthetic models are combined for a realistic simulation of a given phenomenon. In this case, taken a real temporal network and its randomly shuffled variates, synthetic dynamical processes are simulated on the top of them to understand how much structural and temporal correlations, present in the network, influence the unfolding of the synthetic dynamical process. To answer this question, we assume that taken some initial conditions for the process, information (infection, influence, etc.) in the temporal network can pass between

[1] Note, that it is possible to define grand canonical ensembles of temporal networks, but as most of the published methods (including the ones discussed in this Chapter) fall within the micro-canonical definition, we limit our discussion to these cases.

nodes only at the times of their interactions and in a direction respecting the orientation of the actual event (if not noted otherwise). This way the initial conditions and the order of interactions determine the possible time respecting paths, which along the information can be transmitted, and in turn the final outcome of the dynamical process. To carry out such data-driven model experiments we follow general algorithmic recipe:

1. Take a temporal network.
2. Carry out a large number of simulations of a dynamical process on the temporal network with given initial conditions until it reaches a pre-defined equilibrium state.
3. Compute average quantities characterising the dynamics and final outcome of the dynamical process.
4. Apply a selected RRM on the temporal network to eliminate the effects of selected correlations and repeat points 2-3.
5. Compare the average outcome of the dynamical processes simulated on the original network and its RRMs to draw conclusions.

For demonstration [195, 218], as a temporal network we use here a six months long sequence of mobile-phone call interactions (DS1 in Section 1.3.2), and as a synthetic process we use one of the simplest dynamical model, the *Susceptible-Infected* (SI) process, which is to simulate information or infection spreading on networks [43]. This model assigns one of two mutually exclusive states, susceptible (S) or infected (I), to each node in the network and assumes that initially every node are susceptible expect one infected seed, which is chosen randomly at the beginning of the process. During the process, a susceptible node can become infected with a rate β (or probability per unit time step) at the time when it interacts with an infected node[2]. In the coming simulations we use a deterministic SI process with $\beta = 1$ and neglect the direction of events, which means that a susceptible node necessarily becomes infected once interacting with an infected node. We initiate the process by setting infected a single randomly selected node at a randomly selected time, and simulate the process until every node get infected, i.e, when the rate of infected nodes $i(t) = I(t)/N$ reaches 1. Since the seeding time of the SI process is randomly chosen, it is possible that the process reaches the last event in the event sequence before every node gets infected. In this case we apply a *periodic temporal boundary condition*, or in other words, we continue the process from the beginning of the event sequence. This method evidently introduces some invalid time-respecting paths [292] connecting events from the end of the sequence to ones in the beginning, but as we demonstrated, this change marginally the overall dynamics of the process [218]. To obtain a statistical characterisation of the dynamics we measure the average infection rate $\langle i(t) \rangle$ and the $P(t_f)$ distribution of full prevalence time, which fluctuates due to the random initial conditions of the 10^3 independent simulations.

Simulations of the SI process on the original event sequence unfold in a surprisingly slow dynamics and takes on average ~ 700 days to reach every node in the network. This is even more puzzling as the underlying static network structure exhibits a small-world property, thus the average shortest paths between the $\sim 4.5M$ nodes is around $\langle \ell \rangle = 12.31$. Consequently, the spreading process on the temporal network deviates from the ideal structural paths and follow time-respecting paths determined by the ordering of the time-varying interactions. Such slow spreading dynamics is not un-precedential in real life. Spreading of pandemics, electronic viruses, and information, follow their own pathways, which are not necessarily topologically efficient and, could be surprisingly slow, e.g., new infections are reported years after the emergence of a new computer virus or the introduction of an antivirus. The observed slow dynamics may be caused by the combined effects of several structural and temporal characters. First of all, static topological characteristics such as prominent community structure (C) have been shown to give rise to considerable decelerating

[2]Note, that we are going to discuss in details this and other model processes in Section 4.3, where we focus on the unfolding and critical behaviour of various dynamical processes on static and temporal network.

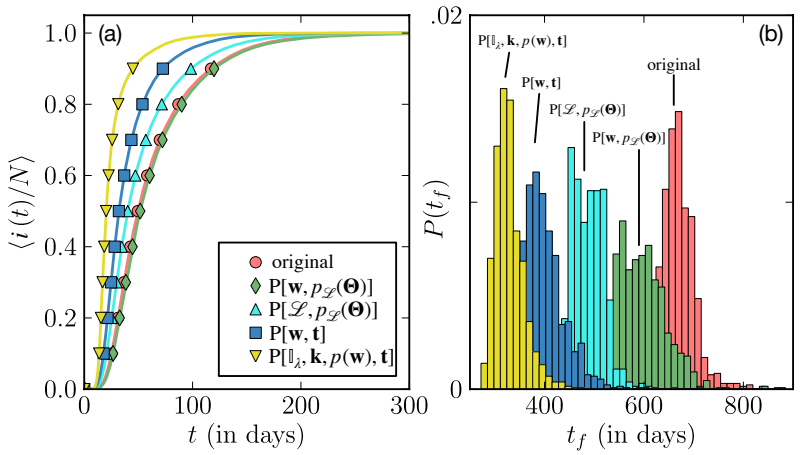


Figure 3.6: (a) Fraction of infected nodes $\langle i(t)/N \rangle$ as a function of time for the original event sequence and null models. (b) Distribution of full prevalence times $P(t_f)$ due to randomness in initial conditions. This figure was published in [195].

effects on spreading speed [295, 114], while a fat-tailed degree distribution has been shown to be an accelerating property [343]. Second, in weighted networks, the relationship between weights and topology (W) provides an additional source of possible influence on the spreading dynamics. In particular for social networks it is known that links within communities are strong, while links between them are weaker [273] - such Granovetterian structure enhances the trapping effect by the communities, leading to further slowing down of spreading [295, 273]. Third, temporal characters like the daily activity cycle (D), the bursty character (B) of interactions and the causal correlations between adjacent interactions (E) (for definition see Section 3.1.4) of the same person may give rise to important temporal inhomogeneities, which in turn influence the speed of spreading dynamics. Next, we introduce some RRMs which remove these effects in a controlled fashion to identify which of them cause the observed slow dynamics.

P[$\mathbf{w}, p_{\mathscr{L}}(\Theta)$] **- Equal-weight link-sequence shuffled model**
Whole event sequences with time stamps (timelines) are randomly exchanged between links that have the same weights, i.e. numbers of events. Timing correlations between adjacent links are destroyed. While temporal characteristics of link event sequences are retained, any correlations between them and the topology are lost. All other temporal and structural correlations are retained. In other words, we constrain our shuffling on the $\mathbf{w}$ number of events (weights) on links, and the $p_{\mathscr{L}}(\Theta)$ distribution of timelines in the static structure $\mathscr{L}$.

P[$\mathscr{L}, p_{\mathscr{L}}(\Theta)$] **- Link-sequence shuffled model**
As above, but sequences are exchanged between links of any weight. Thus, weight-topology correlations are destroyed, while keeping the $\mathscr{L}$ static structure and the $p_{\mathscr{L}}(\Theta)$ distribution of timelines.

P[$\mathbf{w}, \mathbf{t}$] **- Time shuffled model**
The time stamps of the whole event sequence are randomly reshuffled. Thus all temporal correlations with the exception of network-level frequency envelope (for calls, the daily pattern) are destroyed, while the $\mathbf{t}$ event times and topological features like the $\mathbf{w}$ weight sequence are retained.

P[$\mathbb{I}_\lambda, \mathbf{k}, p(\mathbf{w}), \mathbf{t}$] **- Configuration network model**
The original aggregated network is rewired according to the configuration network model, where the degree distribution $\mathbf{k}$ of the nodes and connectedness $\mathbb{I}_\lambda$ are maintained but the topology is

uncorrelated. Then, original single-link event sequences are randomly placed on the links thus the overall weight distribution $p(\mathbf{w})$ remains unchanged, and time shuffling as above is performed keeping the event times $\mathbf{t}$ unaltered. All other temporal correlations, except seasonalities like the daily cycle, are destroyed.

Name	Description	D	C	W	B	E	$\langle t_{20\%} \rangle$
Original	Empirical event sequence	✓	✓	✓	✓	✓	$38.24 \pm .74$
$\mathrm{P}[\mathbf{w}, p_{\mathscr{L}}(\mathbf{\Theta})]$	Equal-weight link-sequence shuffling	✓	✓	✓	✓		$41.20 \pm .83$
$\mathrm{P}[\mathscr{L}, p_{\mathscr{L}}(\mathbf{\Theta})]$	Link-sequence shuffling	✓	✓		✓		$30.67 \pm .55$
$\mathrm{P}[\mathbf{w}, \mathbf{t}]$	Time shuffling	✓	✓	✓			$26.48 \pm .52$
$\mathrm{P}[\mathbb{I}_{\lambda}, \mathbf{k}, p(\mathbf{w}), \mathbf{t}]$	Configuration network model	✓					$17.23 \pm .22$

Table 3.1: Correlations retained in different null models. D: daily pattern, C: community structure, W: weight-topology correlations, B: bursty single-edge dynamics, E: correlations between adjacent events. Average times to reach 20% prevalence with the error of the mean values. This table was published in [195].

In our analysis of the results, first we focus on the two temporal reference models that remove temporal correlations while leaving the static network intact (equal-weight link-sequence shuffled and time shuffled). Fig.3.6a shows that the spreading dynamics for those event sequences that contain bursts (the original and equal-weight link-sequence shuffled sequences) are slower than those for the reference model from which burstiness has been removed (sequences from the time shuffled model). This clearly indicates that burstiness of the event sequences slows down the spreading process significantly. Further, the dynamics for the original sequence and equal-weight link-sequence shuffled model closely resemble each other. This similarity means that event-event correlations have only a small influence on the speed of spreading and it causes even a small acceleration for the process in early times. This can be explained by the emergence of rapid chains of causal events, which helps information spreading locally. However, when looking at the distributions of the full prevalence times t_f, shown in Fig.3.6b, it is seen that for long times, event-event correlations somewhat slow down the process.

Next, we turn to those reference models that modify structural features of the static aggregated network. We find that when the network topology is retained but weight-topology correlations are removed with the link-sequence shuffled reference model, the spreading significantly speeds up compared to the original. This is because the reference model removes the known Granovetterian weight-topology correlations where weak links connecting dense communities of nodes act as bottlenecks. In addition, if topological correlations such as the community structure are removed with the configuration model, the dynamics of spreading becomes even faster.

Finally, we cross-compare the relative importance for the structural and temporal correlations in the call sequence on the spreading speed. As seen in Fig.3.6a, the spreading dynamics for the time-shuffled model where weight-topology correlations are retained but the bursts are destroyed is faster compared to that for the link-sequence shuffled model, where bursts are retained but weight-topology correlations are destroyed. Consequently, burstiness of events on individual links plays a more important role than weight-topology correlations in slowing down the spreading dynamics.

The same conclusions can be drawn once we quantify the slowing down effects by measuring the average times $\langle t_{20\%} \rangle$ to reach 20% prevalence (see Table 3.1). The difference between the original and the fastest model is ~ 21 days, i.e., a factor 2. Similarly for the 100% prevalence this factor also 2 (~ 342 days), showing that the effects of correlations are consistent for the duration of the whole process and for individual runs. As for the effect of the random initial conditions, the small error of mean values in Table 3.1 show that the mean curves in Fig.3.6a characterise the overall behaviour well. The effect of initial conditions are demonstrated in Fig.3.6b, where the

distributions are clearly separable at full prevalence.

As a conclusion we can draw that the bursty non-Poissonian dynamics and the Granovetterian weight-topology correlations are the dominating characters which slow down the spreading dynamics, while other correlations like the community structure or causal event correlations play less important roles. Further, in the related studies [195, 218] we showed that non-Poissonian bursts, evidenced by fat tailed $P(\tau)$ inter-event time distributions, are consistently characterising all activity groups of people. Actually we derived that the decelerating effects of burstiness is coupled with the second moment of $P(\tau)$, leading to the extremely slow dynamics in strongly bursty systems like mobile-phone communication sequences. In addition we found that, interestingly, circadian patterns have negligible effects in influencing the speed of information spreading.

In this Section we briefly introduced the concept and a taxonomy of random reference models and through a data-driven modelling study we demonstrated their potential in characterising temporal networks and ongoing dynamical processes at the system level. As a next challenge we turn our focus on causal correlations between events to see how they lead to the emergence of mezoscopic temporal motifs and to long time-respecting paths in the temporal network.

3.4 Higher order representations

So far we looked at temporal networks as a sequence (or snapshots) of events, which connect two nodes in the network at a given time. Such first order representations have been proven to be useful to characterise temporal structures and ongoing processes but they make it difficult to study the non-Markovian character of temporal networks where events are not necessarily independent from each other. Such correlations, like causal relationships between adjacent events, potentially explain phenomena like bursty interaction patterns [320], temporal motifs [217] and the emergence of long time-respecting paths [217], but can be explanative for the emergence of homophilic motifs of correlated interactions of similar individuals [224]. They can be effectively studied by higher-order representations of temporal structures, where conceptually we identify events (or set of events) as nodes and connect them with a directed link respecting the timely order of the connected events, if they are adjacent at the given order of representation. Such way of description is not only important to study social interactions, but they are central in transportation networks [382], in predicting human mobility [297], or potentially in gene-regulation, or neural networks, just to mention a few examples.

In the following we are going to discuss two methods proposed by us over the last years for higher-order representation of temporal networks. One provides a mesoscopic level description to identify recurrent isomorphic temporal motifs [223, 224], while the other provides a representation at macroscopic scale of the whole temporal structure by mapping all temporal paths simultaneously at once [217]. For simplicity, we assume that nodes in the network cannot participate simultaneously in multiple events, and that events have no duration (although both methods can be easily generalised in this sense). Also note that the investigations of non-dyadic interactions may lead to the recent fields of simplicial complexes [381, 300] and hyper-graphs [185], which are out of the scope of the present thesis.

3.4.1 Temporal motifs

Static motifs are classes of connected isomorphic subgraphs, which appear with a significantly higher frequency in a real structure rather than by chance in a random reference model. Such motifs can be identified as the mesoscopic building blocks of a static network, and more importantly, they can be assigned to special functionality [266]. Analogous definition of motifs in temporal network is a non-trivial problem as it requires the extension of the definition of connectedness, which in case should intuitively include time, and isomorphism, which should respect structural and order

equivalence between sets of adjacent events [223, 224]. As an example, in a social communication network one might detect an event sequence where Alice calls Bob, who then calls Carol and Dave. A similar sequence might be frequently observed to take place between the same people, or between other sets of four individuals. All these sequences are members of the same class, which we call a *temporal motif*. Other example can be found in genetic regulation data, where the event sequence would correspond to regulatory interactions switching on and off as the intercellular system performs its function. Beyond providing insight into the operation of the system under study, temporal motifs allow the study of similarities and differences of temporal networks, as originally proposed for static motifs in [266]. In addition they may help in building models of network evolution [228].

Definition of temporal motifs

Following the logic of static network motifs [266], temporal motifs [223, 224] are defined as temporally connected isomorphic temporal subgraphs of causally correlated events. Their definition relies solely on a sequence of directed interactions $e = (u, v, t)$ to identify causally correlated sets of adjacent events (already introduced in Section 3.1.4), while neglecting any information on the nature/type/content of interactions (which is usually not available anyway). Consequently, their definition is limited to structural and temporal informations to decide about two adjacent events to be causal. One realistic assumption suggests that two adjacent events happening within a Δt time window can be considered causal if the selected time window is sufficiently small. This time window may represent the memory length of a person to remember some information, or the infectious period of an infected agent, or the time one is willing to wait for a connection at an airport. Taking this assumption, we consider two events Δt-*adjacent* if they have at least one node in common and the time difference between the end of the first event and the beginning of the second event is no longer than Δt. Equivalently, two events are Δt-*connected* if there exists a sequence of events $e_i = e_{k_0} e_{k_1} \dots e_{k_n} = e_j$ such that all pairs of consecutive events are Δt-adjacent.

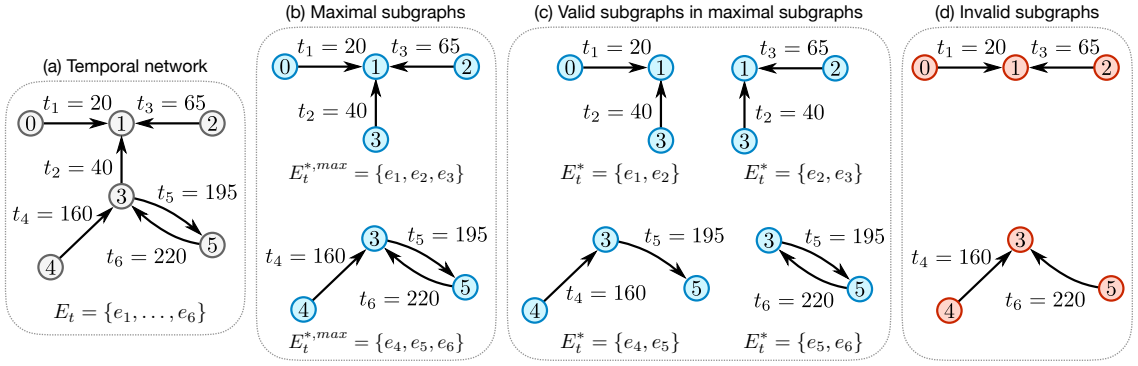


Figure 3.7: (a) An example temporal network with event list E_t of six events. With Δt=40 there are two maximal subgraphs, shown in (b). (c) Valid subgraphs contained in the maximal subgraph in (b). In addition to these the maximal subgraph itself and all unit subgraphs are valid subgraphs. (c) Event sets that are contained in (b) but are not valid subgraphs: the upper one because it does not include all consecutive Δt-connected events, and the lower one because it is not Δt-connected. This paper was published in [224] and partially prepared by L. Kovanen.

Using these definitions, a *connected temporal subgraph* consists of a set of events such that all pairs of events in it are Δt-connected. This ensures that subgraphs are connected both topologically and temporally. While this definition could already be used as a basis for temporal motifs, it suffers from the same shortcoming as its static cousin: in some simple cases the number of connected subgraphs explodes. For example an n-star, where all events take place within Δt, contains $\binom{n}{k}$ connected temporal subgraphs with k events, which would make the resulting motif statistics

difficult to interpret in any intuitive fashion.

One alternative to resolve this shortcoming is to consider only those connected subgraphs where all Δt-connected events of each node are consecutive. This not only solves the problem with the n-star (we now get $n-k+1$ subgraphs with k events) but also offers an intuitive interpretation: each subgraph takes into account all relevant events for each node within the time span covered for that node, in the sense that no events can be skipped (as demonstrated in Fig. 3.7c). We call connected subgraphs that satisfy this requirement *valid temporal subgraphs* and denote them by E_t^*. For every event e_i there is a unique *maximal subgraph* $E_t^{*,max}$ that contains e_i and in which all event pairs are still Δt-connected (as shown in Fig. 3.7b). Note that a maximal subgraph is always a valid subgraph.

Temporal motifs are now defined as classes of isomorphic valid subgraphs, where the isomorphism also includes the similarity of the temporal order of events. Accordingly, two temporal subgraphs are isomorphic if they are topologically equivalent and the order of their events is identical. If a temporal motif is based on a maximal subgraph, we call it a *maximal motif*.

Identification of temporal motifs

Because maximal subgraphs are temporally separated from all other events by at least time Δt, all subgraphs are fully contained in some maximal subgraph. Based on this observation, the process to identify all temporal motifs in a given event set E_t can be separated into three steps:

1. Find all maximal connected subgraphs $E_t^{*,max}$.
2. Find all valid subgraphs $E_t^* \subseteq E_t^{*,max}$.
3. Identify the motif corresponding to E_t^*.

In *step 1*, to find the maximal subgraph which the event e_i belongs to, we start from e_i and iterate forward and backward in time to find all Δt-adjacent events; this process is then repeated recursively with all new events encountered. Assuming the Δt-adjacent events can be found in constant time, the time complexity of this step is $\mathcal{O}(|E_t^{*,max}|)$. Since the same maximal set is discovered starting from any event in it, the total time complexity of this part is $\mathcal{O}(|E_t|)$.

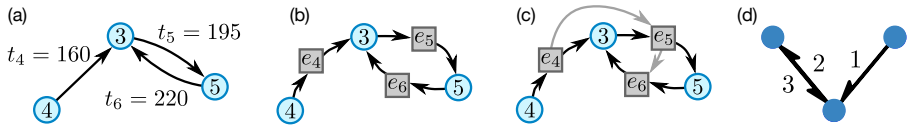


Figure 3.8: Illustration of the algorithm for identifying temporal motifs. (a) A valid subgraph E_t^* with three events. (b) A vertex is created for each event and edges are added to connect them to the corresponding nodes. Colours are used to distinguish between the two types of vertices; the labels of the event vertices are arbitrary. (c) Directed edges are created between event vertices to denote their order. A canonical labelling is then calculated for this graph; all temporal subgraphs that are isomorphic at this stage will yield the same canonical labelling. (d) A concise presentation for the temporal motif. The numbers next to edges denote the order of the events. Note that the numbers are always on the side of the arrow heads. This paper was published in [224] and partially prepared by L. Kovanen

To solve *step 2*, consider the following theorem [223, 224]:

Theorem 3.4.1 Let $G(E_t^{*,max})$ be an undirected graph that has a vertex for each event in $E_t^{*,max}$ and every vertex is connected to the next and previous Δt-adjacent event of both nodes in that event (there are at most four such events). Then every valid subgraph contained in $E_t^{*,max}$ corresponds to a connected subgraph of $G(E_t^{*,max})$.

Proof. Consider a valid subgraph $E_t^* \subseteq E_t^{*,max}$ and the corresponding vertex set in G. Because all event pairs in E_t^* are Δt-connected and the events of every node are consecutive, there is at least

one path between all vertex pairs in this set. Therefore there is at least one connected subgraph of G that corresponds to E_t^*. ■

Now each valid subgraph contained in $E_t^{*,max}$ corresponds to some connected vertex set of G, and the problem of finding all valid temporal subgraphs reduces to identifying all induced subgraphs of G and checking that the events of each node are consecutive. The pseudo-code for identifying vertex sets of induced subgraphs is given in Algorithm 2 (the code assumes that nodes are labeled with integers from 1 to $|V|$). In function FINDINDUCEDSUBGRAPHS we first start $|V|$ search trees so that the tree initialised with node i will include all sets where i is the smallest node. The nodes in the set V_- are excluded from that search tree; initially this set contains all nodes smaller than i. The set V_+ includes the nodes where the search can progress, initially all neighbours larger than i. Because each search tree finds only sets where i is the smallest node, they are necessarily distinct. Now we need to make sure they are complete.

The function SUBFIND first adds the current set to be returned (line 10) and grows then sets recursively. The set V_- is updated to exclude nodes smaller than i; thus each subtree has a different smallest node (from those in V_+) and the subtrees are again distinct. The set V_+ contains nodes where the search may progress: previously allowed nodes larger than i, or neighbours of i not yet excluded.

Algorithm 2 Find the vertex sets of all induced, connected subgraphs of a given graph. The parameter $n_{\max}$ can be used to limit the size of the vertex sets returned. We use $N(i)$ to denote the neighbours of node i.

Require: $G = (V, L)$ is an undirected graph.

```
1:  function FindInducedSubgraphs(G, n_max)
2:      S_all ← ∅
3:      for i in V do
4:          S ← {i}
5:          V_- ← {j ∈ V | j ≤ i}
6:          V_+ ← {j ∈ N(i) | j > i}
7:          SubFind(G, n_max, S_all, S, V_-, V_+)
8:      return S_all

9:  function SubFind(G, n_max, S_all, S_curr, V_-, V_+)
10:     S_all ← S_all ∪ S
11:     if |S| = n_max then return
12:     for i in V_+ do
13:         S* ← S ∪ {i}
14:         V*_- ← V_- ∪ {j ∈ V_+ | j ≤ i}
15:         V*_+ ← {j ∈ V_+ | j > i} \ {j ∈ N(i) | j ∉ V*_-}
16:         SubFind(G, n_max, S_all, S*, V*_-, V*_+)
```

Because the subtrees are distinct at each step, the algorithm will return each set at most once. To see that it returns all possible connected set, consider how we could arrive at an arbitrary connected set S. The search path is rooted at $i_1 = \min S$. Let S_k, $k \leq |S|$, be the set of elements added at depth k. Because S is connected, there is at least one node in $S \backslash S_k$ that is a neighbour of some node in S_k. The only way the construction can fail is if for some k there is a node $i^* \in S \backslash S_k$ that has already been excluded, i.e. it is in V_-. It is not possible that i^* was excluded in the beginning—the tree was rooted at $i_1 \leq i^*$ and only nodes smaller than i_1 were excluded—so it must have happened during the search. But if i^* was added to V_- it means that it was in V_+ but some larger node of S was added

instead, which is a contradiction—in the subtree leading to S we would have added i^*. Hence no node i^* can exist and the construction can always proceed until S is obtained.

Finally for *step 3*, identifying the motif for subgraph E_t^* requires solving the isomorphism problem such that we also include information about the order of the events. One can do this by mapping all relevant information into a directed and coloured graph as illustrated in Fig 3.8, for which the isomorphism can be readily solved with existing algorithms. In practice we calculate for this graph its canonical form, a labelling of vertices that is identical for all isomorphic graphs, so that we can easily tell if two valid subgraphs correspond to the same motif. Finding the canonical form is a non-trivial task, but many efficient algorithms have been developed for this purpose; one is called bliss and is described in [186].

As a final step, to make temporal motifs more accessible we convert the information about the order of events back into plain integers. The above described converting procedure is visually demonstrated in Fig. 3.8 where panel Fig. 3.8d shows a concise presentation of the motif corresponding to the original temporal subgraph in panel Fig. 3.8a.

Temporal motifs in communication networks

To demonstrate our method on a large-scale temporal network, we identify temporal motifs in a mobile-phone call network (a variant of DS1 described in Section 1.3.2), which consists of 320 million mobile phone calls between nearly 9 million customers. We chose the time window to be $\Delta t = 10$ minutes, thus keeping the 35% of events which are Δt-adjacent to at least one other event. As a reference system, to quantify the significance of the observed motifs, we used the time-shuffled RRM ($\mathrm{P}[\mathbf{w}, \mathbf{t}]$ in Section 3.3). Measures in the reference system have been averaged over five independent runs.

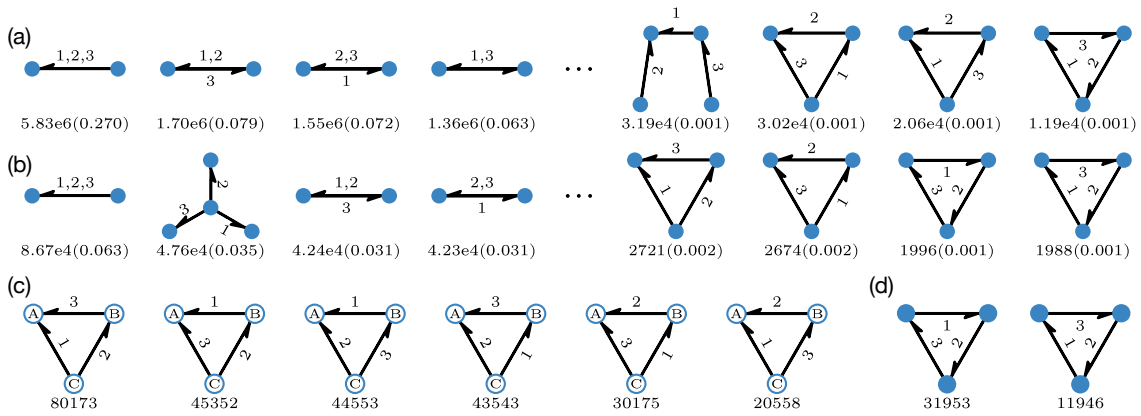


Figure 3.9: (a,b)The four most common (on the left) and least common (on the right) motifs in (a) the empirical data, and (b) unbiased time-shuffled random reference model. The values below each motif denote the total count and, in parenthesis, the fraction out of all motifs with three events. (c,d) The two different kinds of directed triangle motifs with three events. Both groups have been ordered by count in the empirical data that are also shown below the motifs. All motifs in (c), as well as those two in (d), differ only in the order of events. This paper was published in [223] and partially prepared by L. Kovanen

Fig. 3.9a shows the four most and least common three-event temporal motifs (there are 68 three-event motifs in total) in the data, and Fig. 3.9b depicts the same in the time shuffled reference. Unsurprisingly, the number of non-trivial motifs in the reference is lower (only 8.6% of events are Δt-adjacent to some other event) but the two cases still appear qualitatively similar. The most common motifs illustrate the bursty nature of the mobile phone data, while the least common motifs are triangles even though triangles are often considered to be the building blocks of social networks.

The distribution of different motifs is more balanced in the reference: in the empirical data the most common three-event motif makes up 27% of all three-event motifs, but only 6.3% in the time-shuffled reference.

As a further example clarifying this point, we present in Fig. 3.9c and d all motifs based on the different directed triangles with three events. The six motifs in Fig. 3.9c would be equally common in the time-shuffled reference, but in the empirical data we observe a four-fold difference between the most and least common triangles. There are two factors that explain this: burstiness and causality. Burstiness appears in the fact that in the four most common motifs the two calls made by C are consecutive; in the two least common motifs they are not. Causality is most apparent when comparing the most and the least common motifs. In the most common motif the caller of the second call (C) knows about the first call (because he made it herself), and the caller of the third call (B) could know about both previous calls. In the least common motif the caller of the second call (B) cannot know about the first one, and the caller of the third call (C) cannot know about the call made by B. The most common motif is both bursty and causal, while the least common is neither. Causality is also an obvious explanation for the counts in Fig. 3.9d: the triangle where events could cause one another is three times as common as the one where events are independent.

Temporal motifs rely on the higher-order representation of temporal networks as their definition is built on Δt-adjacent events. Although this mesoscopic level characterisation of temporal networks is very informative and useful to identify frequently appearing patterns in the tissue of temporal networks, it is far from providing a complete description of the higher-order structure of the network. In the coming Section we aim to progress in this direction by using adjacent events for a complete and information lossless representation of temporal networks.

3.4.2 Weighted event graphs

Temporal paths in temporal networks are outmost important as they determine the evolving structure of a temporal network and the way any collective phenomena can unfold on its fabric. As we discussed in Section 3.1.4, finding all shortest paths in a temporal network between any nodes at any time is an *NP*-complete problem by using conventional path detection algorithms [91]. A faster way would be to compute paths for a range of values, taking use of redundancy, while even better methods can be designed by using higher-order representations of temporal networks. In this Section, using the concept of event-adjacency, we introduce an information lossless representation of temporal networks. This way of description transforms temporal structures to weighted static graphs, which encodes all temporal and structural informations at once and maps simultaneously all time-respecting paths, with the advantage to be analysed as a static network. Beyond its most general definition, it is capable to constrain on the detection of all Δt-connected temporal paths, which are outmost important in case of limited waiting time process.

Processes with limited waiting times at nodes are particularly sensitive to broad distributions of inter-contact times; the longest inter-contact times may stop the process. Such processes include variants of epidemiological models with recovery mechanism, like Susceptible-Infectious-Recovered (SIR) and Susceptible-Infectious-Susceptible (SIS) models (for definition see Section 4.4.1)[176, 268, 311, 171], where nodes only remain infectious for finite periods. Other examples include social contagion [98, 73], ad-hoc message passing by mobile agents [348], and passenger routing [276]. In these processes, the spreading agent must be transmitted onward from a node within some time Δt or the process stops. This waiting time limit can be directly incorporated into time-respecting paths by requiring that their successive contacts are separated by no more than Δt units of time.

Weighted event graphs [217] are static, weighted, and directed acyclic graphs (DAGs) that encapsulate the complete set of Δt-constrained time-respecting paths for all values of Δt simultaneously. The subset of paths corresponding to a specific value of Δt can be quickly extracted

from the weighted event graph by simply thresholding it. Weighted event graphs can be viewed as a temporal-network extension of the line-graph representation of static networks. There is some similarity with the approach of *Scholtes et al.* [320] that maps two-event sequences onto aggregated second-order networks, and with that of *Mellor* [263] where an unweighted event graph is constructed from pairs of temporally closest events. Our approach [217] builds on concepts introduced in [29, 28].

Definition of weighted event graphs

Let us consider a temporal network $G_t = (V, E_t, T)$ as we defined in Section 3.1.3 (and demonstrated in Fig. 3.10a) with no self-edges or simultaneous events. The weighted event graph representation of G_t is defined as the graph $D = (E_t, E_D, w)$ where the set of nodes E_t is the set of events in G_t and the edges $e_D \in E_D$ represent the adjacency of the events $e_D = e \to e'$ with weights defined as temporal distances $w(e_D) = t' - t$ (see Fig. 3.10b). That is, D is a directed acyclic graph with links weighted with temporal distances, contains all time-respecting paths in G_t. For paths with a waiting time limit Δt, we get the subgraph $D_{\Delta t}$ by thresholding D so that only links with $w \leq \Delta t$ are retained (see Fig. 3.10c).

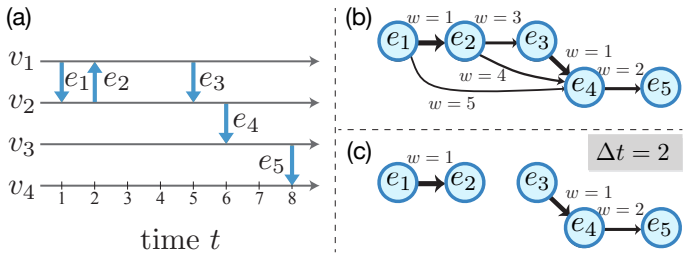


Figure 3.10: Constructing and thresholding the weighted event graph. a) The time line of a temporal network with four nodes $v_1 - v_4$ and five events $e_1 - e_5$. b) The weighted event graph representation of the temporal network. c) The thresholded event graph, containing only pairs of events with a maximum time difference of $\Delta t = 2$. This figure was published in [217] and partially prepared by J. Saramäki.

Constructing the weighted event graph representation $D = (E_t, E_D, w)$ of a temporal network can be done efficiently by noting that the edges in E_D can be listed by inspecting the sequence of events around each node $v \in V$ separately (as described in Algorithm 3). For some data sets the full weighted event graph D might be large, and it is convenient to construct $D_{\Delta t_{\max}}$ that can, for example, be later used to sweep through all values $\Delta t < \Delta t_{\max}$.

For each node in the temporal network $v \in V$ we can build a time-ordered sequence of events $\{e_1, \ldots, e_k\}$ in which v participates. In the case where there are no durations we can then simply iterate over each event e_i, and for each of them search forward in the ordered event sequence until we find an event e_j for which $t_j - t_i \leq \Delta t_{\max}$. We then add a link $e_i \to e_j$ at each step of this process until the event e_j that is too far from the starting event e_i is found. (Note that some Δt adjacencies are found twice.) Creating the event sequences and sorting them can be done in $\mathcal{O}(|E| \log |E|)$ time, and as each step of the algorithm produces a single link (with possibility of some links being visited twice) the algorithm runs in total $\mathcal{O}(|E| \log |E| + |E_D|)$ time. Including the durations of events only requires a small adjustment to this algorithm.

The Δt-thresholded event graph $D_{\Delta t}$ is a superposition of the time-respecting paths that a Δt-limited spreading process can follow. Therefore, its structure tells if the process can percolate the network. A closer look at the problem reveals that here, the concept of percolation is more complex than for static networks. The components of $D_{\Delta t}$ are directed, (even if the events of G are undirected). There are only weakly connected components–there are no strongly connected components because $D_{\Delta t}$ is by definition acyclic. Each event graph node has an in-component and out-component that contain events on up- and downstream temporal paths; these components may

Algorithm 3 Weighted event graph edges for a node [217].

```
function D(v, {e_1, ..., e_k})
    for i ← 1 to k do
        j ← i + 1
        while t_j − t_i ≤ Δt_max and j ≤ k do
            Output: e_i → e_j
            j ← j + 1
```

overlap for different event graph nodes [283]. In the following, we will limit our analysis to weakly connected components because of their uniqueness in $D_{\Delta t}$.

Temporal network percolation

In percolation analysis, the relative size of the largest connected component is defined as the order parameter. Here, there are three ways of measuring the size of a component of $D_{\Delta t}$. (1) One can count the number of *event graph nodes* $S_{E_t}(E') = |E'|$ in a connected component $E' \subseteq E_t$ of $D_{\Delta t}$. This gives an upper bound for the number of events on the time-respecting paths that a spreading process can follow if it includes an event from that component. (2) One can count the number of *temporal-network nodes* $S_G(E') = |\bigcup_{(u,v,t)\in E'}(u \cup v)|$ that are covered by the event graph component E'. This is an upper bound for the number of temporal-network nodes that any spreading process can reach via the component's time-respecting paths. Note that a temporal-network node can belong to multiple event-graph components; this can result in multiple giant components that cover most nodes but are separated in time. (3) One can measure the *lifetime* of the event graph component $S_{LT}(E') = (\max_{(u,v,t)\in E'} t - \min_{(u,v,t)\in E'} t)$. This is an upper bound for the lifetime of any spreading process on the component. Note that there can be many co-existing components with long (or infinite) lifetimes; frequent and sustained contacts between a small number of nodes can already induce such components.

With these measures, we can define the order parameter as the relative size of the largest connected component,

$$\rho_*(D_{\Delta t}) = \frac{1}{N_*} \max_{n_{S_*}\neq 0} S_*, \tag{3.6}$$

where n_{S_*} is the number of components of size S_* for the chosen definition of size $* \in \{E_t, G, LT\}$, and N_* is the maximum possible value that S_* can get as a single component, i.e., $N_E = |E_t|$, $N_G = |V|$, and $N_{LT} = T$. In conventional percolation analysis, the average size of the other connected components is a quantity of interest that is equivalent to magnetic susceptibility. It can be introduced for the $S_*(E')$ event graph components in $D_{\Delta t}$ as

$$\chi_*(D_{\Delta t}) = \frac{1}{N_*} \sum_{S_* < \max S_*} n_{S_*} S_*^2. \tag{3.7}$$

One would expect this quantity to have a maximum at the critical Δt_c, where the percolating connected component emerges in the event graph; in the thermodynamic limit this maximum would become a singularity. However, this quantity might behave differently for $S_G(E')$ and $S_{LT}(E')$ components due to $\sum n_{S_*}$ not being a conserved quantity, and because of the possible multiplicity of giant components in these representations.

Note the link to *directed percolation* [168], where there are two correlation lengths, temporal and spatial, characterising correlations parallel and perpendicular to the directed lattice. In our case, the arrow of time defines the direction, ρ_E gives the probability that a randomly selected event in $D_{\delta t}$ belongs to a structurally percolating infinite cluster, while ρ_{LT} is the typical temporal correlation length for a given δt. However, these correspond to two different order parameters, as the largest and most long-lived components might not be the same.

Weighted event graphs of modelled temporal networks

To explore how Δt controls temporal-network connectivity, we introduce a simple toy model. We define an ensemble of temporal networks $\mathscr{G}_{p,r}(n,k,\lambda)$ where the topology is that of an Erdős-Rényi (E-R) random graph with n nodes and average degree k, and events are generated on each link by a Poisson process with λ events per link on average. We set the observation period T long enough so that $\Delta t \ll T$ and $\lambda \ll T$.

In this model, there is a transition from the disconnected to the connected phase when the independent Poissonian events become Δt-adjacent and form a giant weakly connected component in $D_{\Delta t}$. In terms of degree, a lower bound for this critical point can be estimated as the point where the average out-degree of the event graph becomes $\langle k^{out}_{D_{\Delta t}} \rangle = 1$. In the underlying E-R network, each edge is adjacent to $2(k-1)+1$ edges (including the edge itself), and therefore the average out-degree of $D_{\Delta t}$ is $\langle k^{out}_{D_{\Delta t}} \rangle = \lambda \Delta t\,[2(k-1)+1]$. The condition for the critical point can then be written as

$$k_c = \frac{(\lambda \Delta t)^{-1} - 1}{2} + 1 \quad \text{and} \quad \Delta t_c = \frac{1}{\lambda(2k-1)}. \tag{3.8}$$

This theoretical line $\Delta t_c(k)$ is shown together with simulated results in Fig.3.11a, with the number of events determining the relative size of the largest component, ρ_E. $\Delta t_c(k)$ separates the simulated percolating and non-percolating regimes well. Fig.3.11.b and c show the relative largest component sizes in terms of temporal-network nodes (ρ_G) and component lifetime (ρ_{LT}); a percolation transition appears to take place near the theoretical line $\Delta t_c(k)$ for the number of events from Eq. 3.8. Note generally, the phase transition lines for events, nodes, and lifetime can be different.

Let us investigate the model's critical behaviour in detail, fixing the average degree to $k = 9$. This makes the thresholded event graph $D_{\Delta t}$ dense enough for the mean-field (MF) approach; the MF approximation works well for regular lattices above the critical dimension $d_c = 5$ [168]. We locate the critical point with two methods. First, when the system reaches a stationary state where the order parameter becomes time-invariant beyond fluctuations, the scaling relation $max(S_E) \sim |D_{\Delta t}|^{\beta}$ is expected to hold around the critical point Δt_c, where $|D_{\Delta t}|$ is the size of the thresholded event graph in events, and β is the critical exponent of the order parameter. We measured this relation for several system sizes and values of Δt and found a power-law scaling of $S_E(|D_{\Delta t}|)$ around $\Delta t_c \simeq 0.087$ with the exponent $\beta \simeq 0.75$ (see Fig.3.11d). This point is shown as a circle in Fig.3.11a; it is above the analytical estimate, which provides the lower bound for the critical point. Note that for the directed-percolation university class, the MF solution suggests $\beta_{MF} = 1$.

The second way of determining the critical point is to calculate the ratios

$$r(\Delta t, N) = \rho_E(\Delta t, N)/\rho_E(\Delta t, N/2) \tag{3.9}$$

for varying N [201]. These curves should cross around the critical point Δt_c where $r(\Delta t_c, N) = 2^{-x}$, and x is related to the finite-size scaling exponent. In Fig.3.11g, they indeed cross close to $\Delta t_c \simeq 0.087$ with $r(\Delta t) \simeq 0.82$ suggesting an exponent $x \simeq 0.2863$, should be compared to $\beta/2 \simeq 0.375$.

Finite-size scaling in networks is naturally related to the network volume N (number of nodes) instead of a linear size scale ℓ, which usually cannot be defined. Assuming that $N \leftrightarrow \ell^d$, one can derive finite-size scaling functions, which are expected to hold in the conventional mean-field regime $d > d_c$, or for dense networks. This leads to a finite-size scaling function of the order parameter:

$$\rho_E(\Delta t, N) \sim N^{-\beta/d\nu^*} \widetilde{\rho}_E(N^{1/d\nu^*}(\Delta t - \Delta t_c)), \tag{3.10}$$

where $\nu^* = 2/d$ is the finite-size scaling exponent (of linear size), which depends on the dimension d. If $d < d_c$ it is the spatial correlation length exponent, and above the critical dimension $d_c = 5$ it

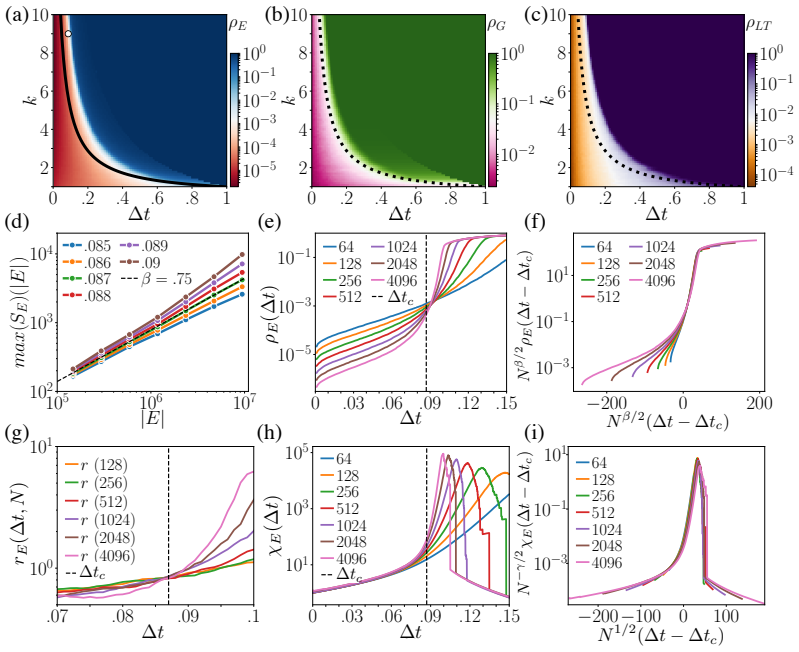


Figure 3.11: Phase diagrams for the random temporal network model as a function of the average network degree k and the maximum waiting time between events, Δt. The colour maps show the (ensemble-averaged) relative size $\rho_*(k,\Delta t)$ of the giant weakly connected components, measured as (a) the number of events in the event graph components S_E, (b) the number of temporal-network nodes that the largest event graph component covers, and (c) the lifetime of the event graph component S_{LT}. The solid line in (a) (dashed lines in (b) and (c)) is the analytic estimate of Eq.3.8. The circle in the upper left corner shows the critical point for $k = 9$ determined as explained in the text. (d) Scaling of $\max(S_E)$, the size of the largest weakly connected component in $D_{\Delta t}$, with the size of $D_{\Delta t}$ measured in number of event-nodes $|D_{\Delta t}| = |E|$, for different Δt. The dashed line assigns the critical $\Delta t_c = 0.87$. (e) The order parameter $\rho_E(\Delta t)$ for different network sizes $N = |V|$. (f) Same as (e) after finite-size scaling using the function defined in Eq.3.10. (g) The ratios $r(\Delta t, N)$ crossing at Δt_c. (h) Susceptibility curves $\chi_E(\Delta t)$ for different sizes. (i) Same as (h) after finite-size scaling using the function defined in Eq.3.11. Dashed lines in labels (i), (g), and (h) show the critical point determined in (d). Computations for (a-c) are with network size $|V| = 2048$ over for $T = 512$ time units with an event rate of $\lambda = 1$ averaged over 10 realisations. Results for (d-i) have the same parameters but are averaged over 100 realisations and may differ in size. This figure was published in [217].

takes the value $\nu^* = 2/d_c$ [201]. At the same time a similar scaling function is expected to hold for susceptibility:

$$\chi_E(\Delta t, N) \sim N^{\gamma/d\nu^*} \widetilde{\chi}_E(N^{1/d\nu^*}(\Delta t - \Delta t_c)), \tag{3.11}$$

where γ is the mean cluster-size exponent. From the definition of χ_E (in Eq.3.7) and the scaling of $\rho(\Delta t, N)$ at Δt_c we can derive the simple exponent relation $\gamma/(d\nu^*) = 1 - \beta/(d\nu^*)$, where $\nu* = 2/d$, $d = d_c = 4$ and $\beta \simeq 0.75$, which gives us a value $\gamma \simeq 1.25$ (which is slightly different from the directed-percolation MF value of $\gamma_{MF} = 1.0$).

To check whether the predicted finite-size scaling behaviour holds around the critical point, we took the simulated $\rho_E(\Delta t, N)$ and $\chi_E(\Delta t, N)$ measured for various N (see Fig.3.11e and h respectively). Using the scaling functions in Eq.3.10 and Eq.3.11 with the determined exponents, we scaled the order parameter and susceptibility as a function of $(\Delta t - \Delta t_c)$. The expected scaling behaviour appears for both quantities close to the critical point (Fig.3.11.f and i).

Finally, we carried out a percolation analysis on three empirical temporal networks, a mobile-phone call networks (DS1 in Section 1.3.2), in a sexual contact network [311], and an air trans-

portation network [12] (detailed results are not reported here but in [217]). This analysis showed that these empirical systems goes through similar percolation transitions as we demonstrated for the toy-model. The percolation point Δt_c is identifiable in each cases, with $\Delta t_c \sim 4$ h 20 min for the calls, $\Delta t_c \sim 7$ d for the sexual-interaction network, and ~ 20 minutes for the transportation network. A spreading agent has to survive at least this long at a node to percolate the network, or in case of transportation, this time-scale corresponds the minimum time required to change flights at an airport. In addition, scale-invariant distributions of structurally and temporally connected components verify the located critical points in each system.

Directed percolation of temporal networks

During the our earlier analysis of weighted event graphs we always considered its weakly connected components. This structure represented the superposition of all connected time-respecting paths in a temporal network, but without taking account precisely the in- and out-components of each node at different times, but operating with their union instead. This way weakly connected components provided us an upper estimate of any macroscopic phenomena emerging on the network and signaled a percolation like phase transition as the function of Δt. Yet this way of description is limited and not appropriate to precisely answer the question, whether temporal networks can be interpreted as directed percolation processes. In a recent set of studies we addressed this challenging question in multiple steps.

First, in [31], we took a complementary approach to the Newman-Ziff algorithm and develop a method to make accurate estimates of the sizes of source and influence sets of every single event in a temporal network, given an arbitrary Δt. We rely on the DAG character of the event graph representation, which allows us to convert our temporal reachability problem to a DAG reachability problem, a.k.a., the graph-theoretical challenge to estimate transitive closure sizes [83]. Relying on already developed probabilistic counting methods [165], we devised an algorithm, which estimates the global reachability for each event even in extremely large temporal networks with hundreds of millions of events. Further, using this approach, we could effectively identify with high probability events with the largest out- (and in-) components in massive temporal networks.

Second, in [33] we studied simple models of temporal networks to show a mapping between their reachability phase transition and directed percolation. Using an event graph reduction technique proposed by Mellor [263], and our out-component size estimation method, we introduced the precise definition of basic thermodynamic quantities for temporal networks. The control parameter, similar to our earlier studies was identified as Δt, the order parameter was measured as the size (in terns if nodes and events) and length of the Δt-limited waiting-time reachability set starting from a random event. Further, in parallel to directed percolation, we defined the survival probability $P(t)$ as the probability of an existing path from a randomly selected initial source event at t_0 to an event after time $t_0 + t$. We also defined particle density $\rho(t)$ as the fraction of infected nodes at time t. Finally, by incorporated the effects of an external field h to this scenario as the spontaneous emergence of sources of infection modeled as an independent Poisson point process. This way we arrived to a more accurate definition of temporal network susceptibility measure $\xi(\tau,h) = \frac{\partial}{\partial h}\rho(\tau,h)$, as the function of τ, indicating distance from the critical point. Relying on a mean-field approach and finite-size scaling techniques, we successfully demonstrated that the examined model temporal network recovers the directed percolation critical exponents (β, $\nu_\parallel$, $\nu_\perp$, and γ) around a phase transition point at Δt_c, mapping it to the directed percolation universality class. In addition, we also observed similar phase transitions in real systems, without the goal of identifying any match of critical exponents.

Third, in [32], we demonstrated that this mapping is robust for a larger set of temporal networks with structural and temporal heterogeneities.

3.5 Generative models of temporal networks

Since the birth of network science, generative modelling is the most important and frequently used modelling technique to understand the emergent properties of complex networks. Such models are overall useful to identify and to test underlying mechanisms, which lead to features characterising real network features. School examples are preferential attachment, which lead to degree heterogeneities, or triadic closure, which can explain the emergence of community structure in networks. However, although simulations of generative models are necessary dynamical, usually the final emergent aggregated structure has been taken as a result and used further as a static structure. Only recently, some new mechanistic modelling techniques have been proposed to simulate temporal networks, where beyond the emergent structural properties, the dynamical features of interactions are also taken into account. It has been recognised, that this finest level of description of networks is necessary for the deeper understanding of emergent global structural and temporal properties, as after all they are the consequences of individual decision mechanisms driving single interactions between people.

In the following we are going to study one promising direction for the generative modelling of temporal networks. We will introduce the basic concept of *activity-driven network models* and will discuss various ways to extend this framework with decision mechanisms, which lead to more and more realistic models of temporal networks.

3.5.1 Activity-driven network model

The modelling framework of *activity-driven networks* (ADN) has been first proposed by Perra et al. for the agent-based simulation of time-varying networks [299]. This model builds on a single assumption that people are not active with the same pace but there are individual differences in the number of interactions one participates due to differences in personality, age, sociability, etc. Such variation has been found in social systems, where communication frequencies were shown to vary from people-to-people over several orders of magnitudes in a larger population. In the ADN framework these variations are introduced a-priory through a quantity called the *activity potential* x_i associated to each N number of nodes in the network with values sampled from an arbitrary distribution $F(x_i)$. Note that $x_i \in [\varepsilon, 1]$ is defined with lower bound assuming a minimum activity level ε. The activity potential is a time-invariant function characterising the activity level of agents by determining their $a_i = \eta x_i$ probability to participate in an interaction per a unit time. Here η is a time rescaling factor assuring the average number of active nodes per unit time to be $\eta \langle x \rangle N$. Built on these definitions, the ADN model is introduced as an iterative process evolving through global time-steps of Δt length, in which on average each node is updated once. More precisely, the generative network process is defined by taking N disconnected vertices at each discrete time step t and activate each vertex with probability $a_i \Delta t$. Once a node is active it generates m links (temporal events) that are connected to randomly selected vertices. Finally, in the end of each iteration step, all links are deleted and the next iteration starts again with a disconnected set of nodes. This algorithm is summarised in Alg.4, where we only save the generated events instead of updating the actual network.

Algorithm 4 Activity-driven network

```
1: function ADN(G = (V, E = ∅, a(i)), T, Δt)
2:     for t ← 1 to T do
3:         for N randomly selected nodes i ∈ V do
4:             if rand() ≤ a_i Δt then
5:                 for m randomly selected nodes j ∈ V \ i do
6:                     Output: event(t, i, j)
```

As Perra et al. explains [299], one can easily realise that the ADN model generates a sequence of random structures. Moreover, if we aggregate these structures over time, the nodes' degree in the aggregated network structure will intuitively depend on their activity potential. Actually there are two mechanisms which can increase the aggregated degree of a node. On one hand, it can be increased by new neighbours contacted by the node. Since at each time step $t \in T$, a node with activity potential a_i selects m other nodes *randomly* for interaction, this out-degree can be computed by making an analogy to the Polya urns problem: it will be equal to the number of different balls extracted from a urn with N balls, performing Tma_i extractions [253]. On the other hand, the degree of a node can be increased by incoming interactions of other active nodes. After some simple calculations (for details see [299]), these two terms can be written as:

$$k_T(i) = k_T^{out}(i) + k_T^{in}(i) = N(1 - e^{-Tma_i/N}) + Tm\eta \langle x \rangle e^{-Tma_i/N} \sim N(1 - e^{-Tm\eta x_i/N}) \quad (3.12)$$

in the limit of large N and small T/N. Based on this calculation and by using some further approximation we can show that the emerging degree distributions for early times, where k/N is small, appears as

$$P_T(k) \sim \frac{1}{Tm\eta} F\left[\frac{k}{Tm\eta}\right], \quad (3.13)$$

which indicates that the functional form of the aggregated degree distribution will emerge with the same scaling form as the activity distribution.

This model is very simple and far from being realistic as it produces a sequence of random structures, with an emerging degree distribution trivially the consequence of a parameter of the model. In its simplest form, it does not reproduce common emerging properties of real social networks like link weight heterogeneities, community structure, weight-topological correlations, or burstiness. On the other hand it has the potential to serve as the reference model and the starting point for more realistic network models with integrated microscopic decision mechanisms inducing more realistic network or dynamical characters. As we will see later, there are miriad ways to extend the general ADN model and to use it for hypothesis testing on emerging network features or their effects on dynamical processes. At the same time, ADNs model temporal networks, thus they provide an ideal way to study the importance of multiple temporal scales [310], and dynamical or structural features of interactions on ongoing diffusion, epidemics, opinion formation, complex contagion, etc. processes, just to mention a few examples. I had contributions to extend the general ADN models in two directions, whether by addressing various structural mechanisms [206, 351, 233, 344], or by mechanisms driving the dynamics of dyadic interactions [352]. In the following Section, I will summarise some of these studies to familiarise the reader with the concept and potential of this modelling framework.

3.5.2 Memory processes in egocentric network formation

In the activity-driven framework we have considered only memoryless generative processes so far. At each time step, nodes select their partners with a uniform probability. The model thus neglects the heterogeneous nature of individuals' social interactions, which is a common character of real social systems where egos may have strong and weak ties. The heterogeneity of social ties is a key ingredient of the social structure and plays a crucial role on diffusion processes [288]. Our goal here is to understand the mechanism driving their formation, explicitly considering the network's time-varying nature.

In our analysis [206], we focus on a prototypical large-scale mobile phone call network (DS1 in Section 1.3.2), which structure is characterised by heavy-tailed degree k, link weight w - defined here as the number of call between people, and activity rate a distributions - defined as the probability of any given node to be involved in an interaction at each unit time (see Fig. 3.12a-c respectively).

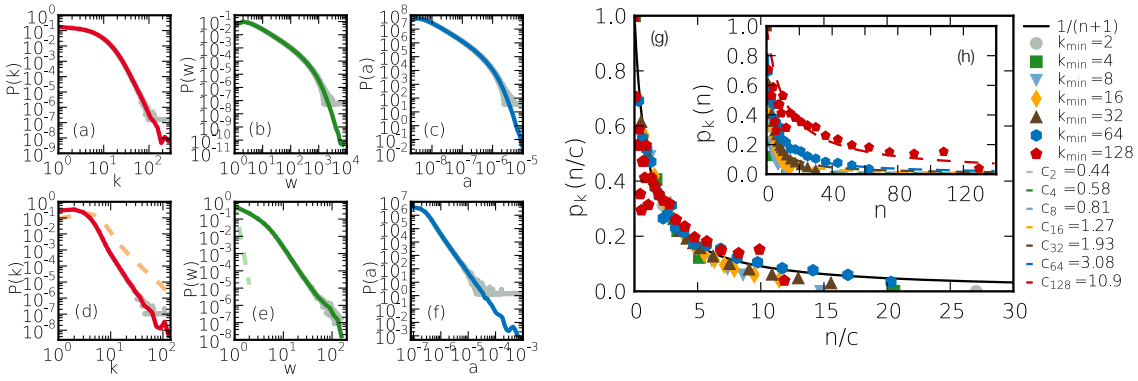


Figure 3.12: Panels (a), and (d) plot the degree distributions; panels (b), and (e) plot the weight distributions; and panels (c), and (f) plot the activity distributions of the empirical and modelled networks (respectively). Grey symbols are the original while coloured symbols are the corresponding logarithmic binned distributions. In panels (d-f) solid lines assign distributions induced by the reinforced process, while dashed lines are due to the memoryless process. Model calculations were parameterised as $N = 10^6$, $\varepsilon = 10^{-4}$ and $T = 10^4$. Panel (g) plots the $p_k(n)$ probability functions calculated for different degree groups in the MPC network. In the inset (h), symbols show the averaged $p_k(n)$ for groups of nodes with degrees between the corresponding $k_{min}...k^2_{min}-1$ values. Continuous lines are the fitted functions of Eq.4.3 with c parameter values showed in the legend. The main panel depicts the same functions after rescaling them using Eq.3.15. The continuous line describes the analytical curve of Eq.3.15. This figure was published in [206].

Our goal first is to identify the mechanisms driving the dynamics of single interactions of the egocentric networks (egonets) and which is responsible for the emergent heterogeneities. We will demonstrate that memory processes characterising each agent plays a crucial role here, which can be incorporated in the ADN model via a simple non-Markovian reinforcing mechanism, that allows to reproduce with great accuracy the empirical data.

Egocentric network dynamics.

As we briefly discussed in Section 1.2.2, social networks are characterised by two types of links. The first class describes strong ties that identify time repeated and frequent interactions among specific couples of agents. The second class characterises weak ties among agents that are activated only occasionally. It is natural to assume that when observing interactions, strong ties are the first to appear in the system, while weak ties are incrementally added to the egonet of each agent, as it has been demonstrated in [226]. To quantify this hypothesis in observations, we observe the interaction (call) dynamics of egos and measure the probability, $p(n)$, that their next communication event will establish a new $(n+1)^{th}$ link, or will be a repeated interaction on one of their n already observed social ties[3]. We calculate these probabilities in the MPC dataset averaging them for users with the same degree k at the end of the observation time. We therefore measure the quantity $p_k(n)$ for the egonets with the same degree k and $n \leq k$. The empirical $p_k(n)$ functions for different degree groups are shown in Fig.3.12g inset (coloured symbols). Interestingly, the probabilities are decreasing with n for each degree class indicating a slowing down in the egocentric network evolution. The larger the egocentric network, the smaller the probability that the next communication will be with someone who was not contacted before. In other words, agents have memory to remember their social ties and they tend to repeat interactions on them.

The empirical growth of the egonet can be captured by a simple mechanism. We find that the probability that a node, characterised by a social circle of size n, will establish a new tie is well

[3]Note, that here and later in the Section, n is equivalent with the degree k of a node at a given time t. We distinguish between these notation to avoid confusion between the static degree k (here when $t = T$) and the evolving degree n (at time t) of a node.

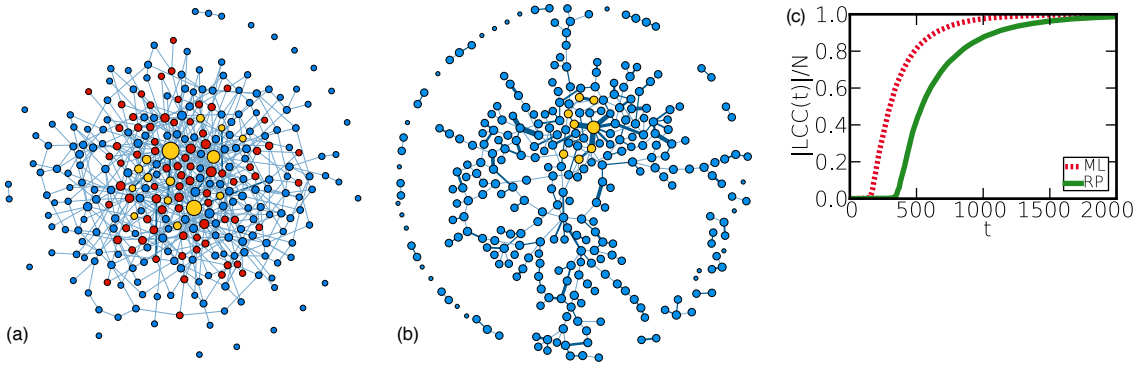


Figure 3.13: Emergent structure and rumour spreading processes in (a) ML and (b) RP activity-driven networks. Node colours describe their states as ignorant (blue), spreader (red) and stifler (yellow). Node sizes, colour, and width of edges represent the corresponding degrees and weights. The parameters of the simulations are the same for the two processes: $N = 300$, $T = 900$, $\lambda = 1.0$, and $\alpha = 0.6$. The processes were initiated from a single seed with maximum strength. In panel (c) we show the sizes of the largest connected components (LCC) as a function of time for time aggregated ML and RP networks. Simulations were run with the same parameters considering $N = 10^5$ nodes. This figure was published in [206].

fitted by the expression :

$$p(n) = 1 - \frac{n}{n+c} = \frac{c}{n+c}. \tag{3.14}$$

Analogously, the probability of having an interaction with someone who is already in the egocentric network is $n/(n+c)$. Here c is an offset constant depending on the degree class considered. By fitting the function in Eq.4.3 on the empirical data (solid lines in Fig.3.13g inset) we can determine the corresponding constant c for each degree group and using them rescale the empirical $p_k(n)$ functions as

$$p_k(n/c) = 1/(n/c+1). \tag{3.15}$$

This rescaling collapse the data points of different degree groups on a single curve (see Fig.3.12g main panel), which suggests that the same mechanism is driving the evolution of the egonets of all individuals independently of their final number of connections.

Next question, how can we integrate such a mechanism into the ADN model to test its effect on the emerging structure. Inspired by the observations in the MCP dataset, we impose a reinforcement mechanism in the ADN model definition by extending it in two ways. First, we assume that nodes have memory, in other words they remember the set of other nodes they have been in contact during the course of the simulation. Second, we assume that when a node, with n previously established social ties, becomes active in an iteration step it can make two kinds of actions: (a) it will contact randomly a new node with probability $p(n) = c/(n+c)$, or (b) with probability $1 - p(n) = n/(n+c)$ it will interact with a node already contacted, thus reinforcing earlier established social ties. In this case, the selection is done randomly among the n actual neighbours. This model, that in the following we will denote as RP (reinforcement process), is non-Markovian as memory is explicitly introduced in the ego network dynamics. For the moment we fix $c = 1$ for all the nodes while we will generalise our model later [351].

A side by side comparison of the time-aggregated representations of networks generated by the memoryless (ML) and reinforced process (RP) models (using the same parameters) is shown in Fig.3.13-a and b (where colours should be disregarded for the moment). In both cases we assumed that the distribution of the activity potential followed a power-law $F(x) \sim x^{-\nu}$, with an exponent value $\nu = 2.8$ matching the observed MPC value. As a consequence, in case of the ML

dynamics (Fig.3.13a), we obtained an aggregated network with a degree distribution $P(k) \propto k^{-\gamma}$ where $\gamma = \nu$ and a weight distribution decaying exponentially [299, 330]. This is also confirmed by large scale simulation results reported in Fig.3.12d and e (dashed lines). In case of the RP dynamics (Fig.3.13.b), the memory process induces a considerably different structure. These effects are quantified in Fig.3.12d, e, and f (solid lines). We observe a degree distribution that is heavy-tailed but more skewed in the RP model than the ML ($\nu_{RP} > \nu_{ML}$). This distribution is qualitatively matching better the corresponding empirical measure in Fig.3.12a. Furthermore, the RP model generates heterogeneous weight distributions (see Fig.3.12e solid line) capturing extremely well real data (see Fig.3.12b). This is not the case in the ML model where the absence of memory induces exponential weight distributions far from reality (see Fig.3.12e dashed line). The RP dynamics not only induces realistic heterogeneities in the network structure, but also controls the evolution of the macroscopic network components. Indeed, due to the reinforcement mechanism, the largest connected component (LCC) in RP networks grows considerably slower than in the case of ML models (for illustration see Fig.3.13.c). This is an important feature because dynamical process evolving on time-varying networks will progress with a time-scale that cannot be smaller than the LCC growth time-scale. As consequence, any dynamical phenomena taking place on time-varying networks with memory will evolve at a slower rate than in memoryless time-varying networks.

3.5.3 Individual heterogeneities in social capacity

In the definition of the memory driven ADN model we took some assumptions for simplicity, which limited our model to account for differences potentially characterising individuals. First of all we assumed that the scaling constant in Eq. 4.3 is a constant value for everyone. This constant determine the intrinsic characteristic limit of an ego to maintain multiple ties, thus it may vary from a person to another. At the same time we assumed that speed of exploration of new friends is entirely coded in the activity potential of people. Recent findings demonstrate, that this ability may vary strongly between people depending whether they are more like *social explorers* or *social keepers* [267]. Such variance can be easily incorporated in our memory function as:

$$p_b(n) = \left(1 + \frac{n}{c_b}\right)^{-\beta_b} \tag{3.16}$$

where β_b modulates the tendency to explore new connections. Just as earlier, here we assign nodes to degree classes b containing actors with statistically equivalent characteristics, i.e. nodes that engaged a similar number of interactions and that feature a comparable cumulative degree in the observation period. We have shown in [351] that this functional form fits very well on memory functions measured in several empirical temporal networks like Twitter mention interactions (TWT), scientific co-publications (PRA, PRB, PRL), or in case of mobile phone call (MPC) interactions. Moreover, we found that the β_b parameter of the $p_b(n)$ functions appears highly invariant over various final degree groups. One exception was the MPC network, where some variance and negative correlations have been found.

By leveraging on this result we can define a variation of the memory driven ADN model by implementing $p_b(n)$ in the decision process. Moreover, for this model it is possible to write explicitly the Master Equation (ME) describing the evolution of the probability distribution $P_i(n,t)$ that a node i has degree n at time t:

$$P_i(n,t+1) = P_i(n-1,t)\left[a_i p_i(n-1) + \sum_{j\sim i} a_j \sum_{n_j} \frac{p_j(n_j)}{(N-n_j)} P_j(n_j,t)\right] +$$
$$P_i(n,t)\left[a_i[1-p_i(n)] + \sum_{j\sim i} a_j \sum_{n_j}\left(1 - \frac{p_j(n_j)}{(N-n_j)}\right) P_j(n_j,t)\right] + P_i(n,t)\left[1 - \sum_j a_j\right]. \quad (3.17)$$

In the above equation the sums in $j \sim i$ and $j \nsim i$ run over the nodes already contacted and not yet contacted by i, respectively. n_j describes the degree of each node j. Moreover, we work in the $a \ll 1$ limit, so that we assume that only one node is active for each evolution step. The first two terms on the right hand side of Eq.3.17 account for the increment of the number of nodes having degree $n-1$. The former occurs when node i having degree $n-1$ gets active and contacts a new node with probability $a_i p_i(n)$. On the other hand the latter one is effective when node i gets contacted by node j of degree n_j (that never got in contact with i before) that activates and attaches to node i with probability $a_j p_j(n_j)/(N-n_j)$. In the latter, the $1/(N-n_j)$ factor accounts for the probability of j to exactly select node i amongst the $N-n_j$ nodes outside of the j's neighbourhood of size n_j. Likewise, the third and fourth terms of the r.h.s. of the equation account for the conservation of the number of nodes of degree n. This is achieved either when node i gets active and contacts one of its neighbours with probability $a_i(1-p_i(n))$, or when i gets contacted by one of its neighbours. The last term of Eq. 3.17 accounts for the possibility in which no node gets active in the current evolution time step, thus conserving the $P_i(n,t)$. Given the $a \ll 1$ approximation this term reads $\prod_j (1-a_i) \simeq 1 - \sum_j a_j$.

In case the network is characterised by a single exponent β, in the large time and degree $1 \ll n \ll N$ limits (so that n can be approximated by a continuous variable and $N-k \approx N$), the probability distribution $P_i(n-1,t)$ can be obtained explicitly as

$$P_i(n,t) = A \exp\left[-\frac{\left(n - B(a_i,c_i)t^{\frac{1}{1+\beta}}\right)^2}{Ct^{\frac{1}{1+\beta}}}\right], \quad (3.18)$$

where A is a normalisation constant, C a constant and $B(a_i,c_i)$ a multiplicative factor of the $t^{1/(1+\beta)}$ term that depends on the activity a_i and c_i of the considered agent a_i (for further details see [351]). From Eq. 3.18 we can determine the evolution in time of the average degree $\langle n(a,t)\rangle$ of nodes belonging to a given activity class as:

$$\langle n(a,t)\rangle \propto (at)^{\frac{1}{1+\beta}}. \quad (3.19)$$

The growth of the system is thus modulated by the parameter β that sets the strength of the reinforcement of ties. The validity of Eq. 3.19 is demonstrate in Fig. 3.14a-d, where the average degree evolution is shown for various activity groups in different empirical datasets, together with analytical predictions. Note that in the limit $\beta = 0$ the growth would be linear and in the opposite limit $\beta \to \infty$ each node would create, and constantly reinforce, just one tie, i.e. the first established. Furthermore, Eq. (3.19) connects, at a given time t, the actual degree n and the activity a of a given node, as $n \propto a^{\frac{1}{1+\beta}}$. Thus, given any specific activity distribution $F(a)$, we can infer the functional form of the degree distribution $P(n)$ by substituting $a \to n^{1+\beta}$, finding:

$$P(n)dn \propto F(n^{(1+\beta)})n^{\beta}dn. \quad (3.20)$$

This prediction is demonstrated on empirical data in Fig. 3.14e-h. It is important stressing that the analytical framework is not limited to a specific functional form of the activity. Indeed, with an

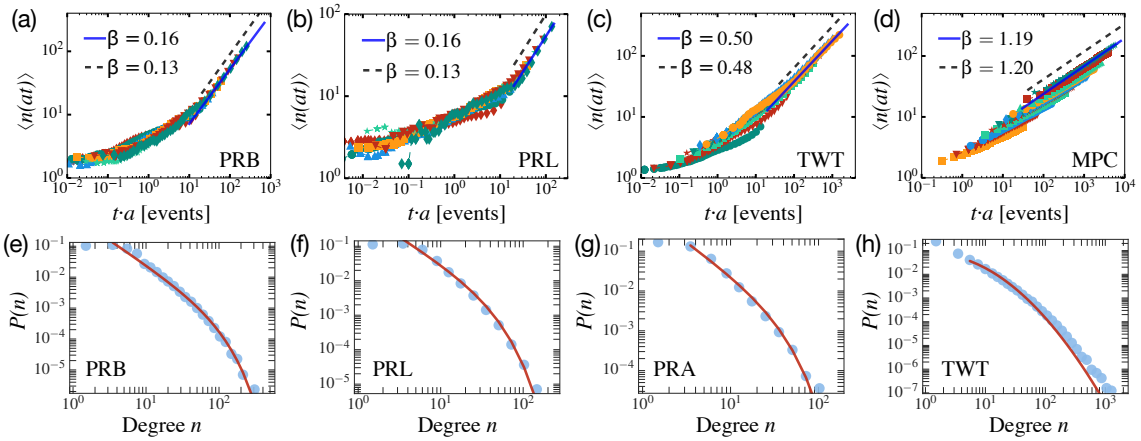


Figure 3.14: The rescaled $\langle n(at)\rangle$ curves for selected nodes classes belonging to the (a) PRB , (b) PRL, (c) TWT, and (d) MPN datasets. The time of the original data (symbols) is rescaled with the activity value $t \to at$. We also show the fitting curve $\langle n(t)\rangle \propto t^{1/1+\beta}$ (blue solid lines) and the expected asymptotic behaviour (black dashed lines). In the MPN case (d) we fit using $\beta = \beta_{\min} = 1.2$. (e-h) The degree distribution $P(n)$ for the (e) PRB, (f) PRL, (g) PRA, and (h) TMN datasets. The predicted functional form of $P(n)$ found in Eq. (3.20) is shown for comparison (red solid lines). We show the data starting from the lower bound of the degree distribution, including in the plot all the statistically significant measures of the probability density function $P(n)$. This figure was prepared by E. Ubaldi and published in [351].

arbitrary functional form of $F(a)$, Eq. (3.19) gives us the possibility to predict the behaviour and parameters of the corresponding degree distribution (for demonstration see [351]).

In case of different values of β in the system complicates the model beyond analytical tractability. Nevertheless, we find that the leading term of the evolving average degree can be described by introducing a simplified model, where the minimum β value observed in the ensemble [351].

3.5.4 Dyadic closure and node removal mechanisms

Building on the model with the simple memory process we discussed in Section 3.5.2, here we further introduce three mechanisms, which are assumed to shape social networks [233]. They arguably lay behind the emergence of several realistic characters like the community structure, weight-topological correlations, and the stationary evolution of the temporal network structure.

We consider two *dyadic closure* mechanisms [233], one called *cyclic closure*, responsible for triangle formation in social networks, it shapes the social structure at mesoscopic scale, and it leads to the emergence of communities [162]. The other mechanism called *focal closure*, on the other hand, is independent of network structure and represents the formation of ties between individuals with shared attributes or interests. It is driven by the propensity to seek cognitive balance between connected egos [164, 155] as suggested by earlier theories in sociology [323, 308]. An applicative definition of *cyclic* and *focal closure* in general is given by Kumpula *et al.* [228, 227], who modelled cyclic closure as biased local search, as contrary to *focal closure*, which is modelled as an unbiased global random search. Finally, the third ingredient of the model is a *node removal* process, which ensures the network to reach an equilibrium state where its overall characteristics become invariant of time.

We introduce these mechanisms in the memory-driven ADN model by first, at each iteration step, deleting nodes with probability p_d. To keep the network size constant, for each deleted node, we add a new disconnected node to the network in the next iteration step. If a node is not deleted, we let it to follow the memory driven activation process but when it repeats an interaction on an existing link, instead of completely randomly, it selects one of its neighbours j with probability $p^w_{ij} = w^t_{ij} / \sum_{k \in V^i_t} w^t_{ik}$ weighted by the number of their past interactions. The two nodes then interact

and increase their link weight w_{ij}^t by δ, a parameter which mimics a *social reinforcement process*. On the other hand, if the node decides to form a new link, it may follow different strategies. In all cases, the new tie will initially have unit weight $w^t = 1$. If the degree of the focal node is 0, it randomly picks another node from the entire network j (*focal closure*) and forms a tie. Otherwise, it attempts to create a new link with a triadic closure mechanism. First, it chooses one of its neighbours j randomly with a weighted probability p_{ij}^w. If j has no other neighbours than i, node i looks for another random node to interact with (*focal closure*) and forms a link. Otherwise, it looks for a random neighbour k of j ($i \neq k$) with a weighted probability p_{jk}^w. If k is not an already existing neighbour of i ($k \notin V_t^i$), the two nodes interact with probability p_Δ, and close the triad by forming a link (*cyclic closure*). Otherwise, with probability $1 - p_\Delta$, node i follows the *focal closure* strategy and instead forms a link with a randomly selected node (other than j and k). Finally, if k is already a neighbour of i, that is $k \in V_t^i$, the two nodes interact and increase the weight of their existing link by δ (*reinforcement process*). At the end of each iteration step, all nodes finish their active interactions but remember their already connected neighbours $j \in V_t^i$ and the weight w_{ij}^t of interactions with each of them. For a pseudocode of the algorithm see [233].

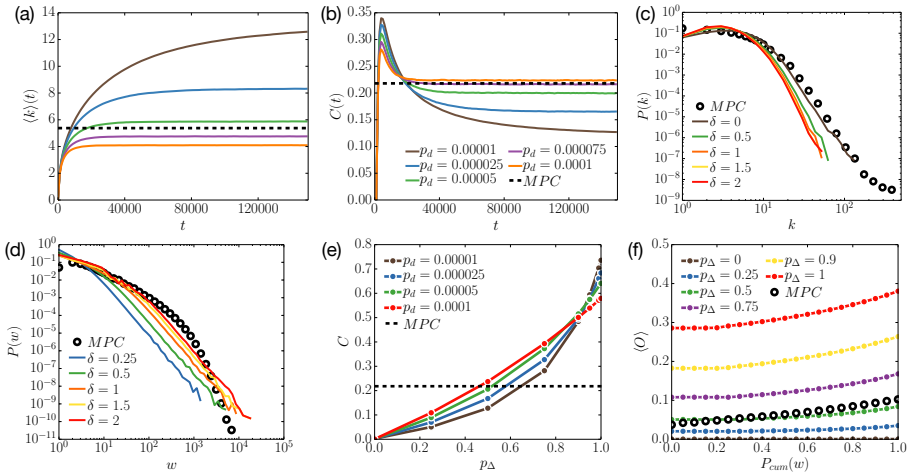


Figure 3.15: The evolution of (a) the average degree $\langle k \rangle(t)$ and (b) clustering coefficient C as the function of time and deletion probability p_d (for exact values, see legend). Panel (c) and (d) depicts the $P(k)$ degree and $P(w)$ weight distributions (respectively) of model networks with varying δ values. Panel (e) shows the dependence of the average local clustering coefficient C on the parameter δ, while emerging weight-topology correlations are shown in panel (f) as the function of the same parameter measured as the average overlap $\langle O \rangle$ vs the cumulative tie strength $P_{sum}(w)$ in modelled networks. Black circles and dashed lines denote the corresponding empirical MPC measures. This figure was published in [233].

In addition to the activity-driven model parameters whose values are fixed ($\eta = 1$, $\varepsilon = 10^{-3}$, $\Delta t = 1$ and $\gamma = 2.8$), our model has three intensive parameters, p_Δ, p_d, and δ. By varying them, one can simulate a rich variety of time-varying networks with several emergent structural properties and correlations. In the following, we explore how the properties of the emerging network structure depend on time and on the intensive parameters, and whether these properties match empirical observations. As a real reference network we use an aggregated representation of a mobile phone communication network (DS1 in Section 1.3.2), with characters shown as black dashed lines or empty circles in Fig. 3.15. In the following, model networks were generated via simulations with $N = 10,000$ nodes (if not noted otherwise), and results were averaged over 100 independent realisations. We measured network characteristics by considering links that are actually present in the network, i.e. we disregarded links of removed nodes.

The simulated networks are inherently temporal thus they generate time-varying interactions

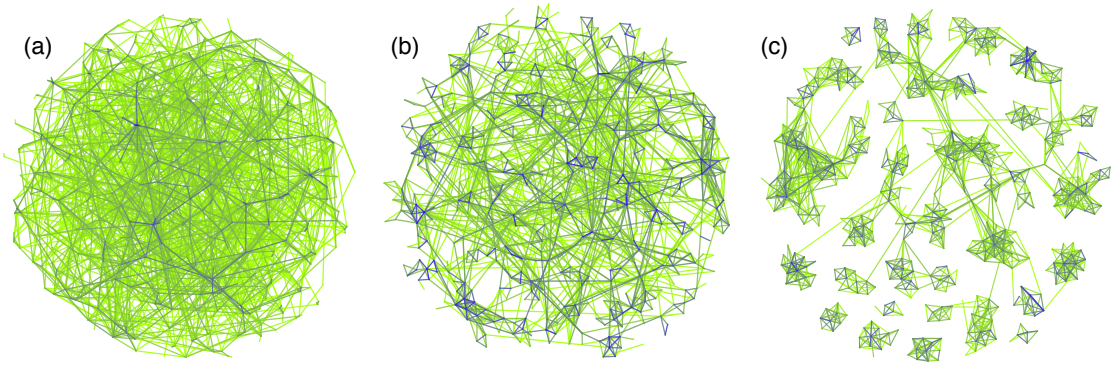


Figure 3.16: Demonstration of the emerging structure in the time-varying network model. Panels (a-c) depict simulated networks with fixed $\delta = 1$ and varying $p_\Delta = 0.5$, 0.9, and 0.995 ($p_d = 4e-5$, $2e-5$, and $1.04e-5$) respectively. Each panel depicts the actual structure of a network with $N = 500$, in its stationary state. Links are coloured according to their weight (darker link colour = stronger link weight). This figure was published in [233].

and an evolving network structure. To explore their evolution, we measured their $\langle k \rangle(t)$ average degree and the $C(t)$ average local clustering coefficient as functions of time. Since each process starts from a set of disconnected agents, and hence all measured properties are trivially zero at time t = 0. However, as time goes by and ties are formed via temporal interactions, $\langle k \rangle(t)$ increases until the network reaches a stationary state with constant $\langle k \rangle(t)$. The time which takes to reach this equilibrium state strongly depends on the node deletion probability p_d as shown in Fig. 3.15a. The same is true for the clustering coefficient (see Fig. 3.15b), however before its equilibrium, it increases due to triangles introduced in the beginning of the process, and then it decreases once ties are formed and new interactions appear only as reinforcement or by creating random links. Such ties decrease the relative density of triangles until the clustering coefficient reaches a stationary value.

Next we see how the emerging structure depends on the reinforcement parameter δ. We fix $p_\Delta = 1$ and $p_d = 5 \times 10^{-5}$ and vary δ between 0 and 2, while measuring network properties in the stationary state. As we see in Fig. 3.15c and d the network emerges with heterogeneous degree and weight distributions, which relatively well approximate the empirical distributions (shown with black empty circles). They appear with a tail robust against δ but they are shifted, especially $P(w)$, assigning that the average weight naturally depends on the strength of reinforcement.

At the same time cyclic closure mechanism dominantly control the emergence of triangles and cluster in the network. This is shown in Fig. 3.15e (where we set $\delta = 1$), where by increasing p_Δ, cyclic closure becomes more dominant, reflected by the steep increase of clustering. Several of the depicted $C(p_\Delta)$ curves cross the empirical value, indicating that various sets of parameters (p_Δ and p_d as well) are eligible to yield networks with realistic clustering value. In addition, we measured weight-topology correlations to check if the emerging model networks recover the Granovetterian weak-tie structure (for more on this theorem see Section 1.2.2). More precisely, we computed the average link overlap $\langle O \rangle$ (for definition see Section 1.2.1) as the function of the cumulative tie strength $P_{cum}(w)$ [288]. In the MPC network, this quantities are positively correlated (black circles in Fig.3.15f) in accordance with earlier observations [288]. Interestingly, such correlations spontaneously emerges in the modelled networks for any positive values of p_Δ (see Fig. 3.16f) or δ (not shown here [233]). For larger p_Δ, this function remains positive and shifted upward as more triangles evolve. More importantly, the combined effects of these two mechanisms, cyclic closure and reinforcement, lead to the emergence of a community like structure in the modelled aggregated network. It is demonstrated in Fig. 3.16a-c where we plot aggregated structures for increasing p_Δ

(left to right). When p_Δ is small, the network structure is like a densely connected random structure, but with heterogeneous weights due to the present reinforcement process ($\delta = 1$). On the other hand, by increasing p_Δ, communities emerge simultaneously with strong ties inside communities, and weak ties between them (darker link colour = stronger link weight). The cyclic closure alone would not be eligible for the emergence of such communities and the Granovetterian picture, as although the clustering of the network wold be large, no strong ties would bias the link creation process and the network would appear rather homogeneous on the mesoscopic scale [233].

* * *

Demonstrated by studies summarised above, activity driven network models have a large potential to define mechanistic models of temporal networks with realistic emerging features. They provide an ideal testbed for identifying the role of microscopic mechanisms in the network formation, and their effects on the final outcome of dynamical processes. This summary already demonstrated the capacities of this framework, thus for the sake of concise description, I decided to excluded two published works of mine from the discussion.

One [344] addresses the role of endogenous and exogenous link creation mechanisms in cross-country and cross-sectoral R&D company alliance networks. The extended ADN model is able to reproduce a number of micro-level measures, including the degree distributions, local clustering, path length, component sizes, and the emergence of alliance clusters. Furthermore, by estimating the link probabilities towards newcomers and established firms from the data, we find that endogenous mechanisms are predominant over the exogenous ones in the network formation, thus quantifying the importance of existing structures in selecting partner firms.

The other study [352] extends the ADN framework by considering strategies what individuals adopt when selecting between new or old social ties (similar to the model described in Section 3.5.3), and the bursty nature of the social activity setting the pace of these choices. In this paper we studied the non-trivial interplay of these two, structural and dynamical, characters and found that the effects of burstiness might be suppressed when individuals exhibit a strong preference towards previously activated ties. These modelling results provides a principled method to classify the temporal features of real networks, and thus yield new insights to elucidate the effects of social dynamics on spreading processes.

3.6 Conclusions

In this Chapter we focused on my contributions to the theoretical and methodological foundation of the field of temporal networks. Beyond the discussion of my overall view and interpretation of this way of description, we walked through several results addressing the representation, macro-, meso- and microscopic level description of temporal networks, while we also introduced several techniques for their data-driven analysis and modelling, applying random-reference and mechanistic modelling paradigms. For the sake of completeness it worth to mention our recent work on temporal network embedding [346], which aimed to identify network nodes and times with similar outgoing reachability sets, thus potentially inducing similar epidemic outcomes seeded from them.

In this chapters, however, we have not addressed another important aspects of (static or time-varying) networks: their effects on ongoing dynamical processes. Such questions will be central in the coming Chapter where we will study collective social phenomena and aim to understand how structural and dynamical characters of networks influence the unfolding of processes modelling social or epidemic spreading.

4. Collective phenomena on networks

4.1 Introduction

Collective phenomena emerge in various ways in complex systems. Assuming a network description, they can appear as cooperative patterns in the structure or the dynamics of interactions, as we just discussed in the previous two Chapters. On the other hand, cooperative phenomena may appear as collective states of nodes in a network. Such patterns may strongly depend on the underpinning structure or could even alter the network formation, this way leading to correlated patterns between the states and interactions of agents. In this Chapter, we are going to focus on the observations and modelling of such phenomena emerging *on networks*.

During this discussion we are going to move along several dimensions characterising collective processes. First of all, one dimension may consider the coupling between the phenomena and the underlying network. If the system exhibits *one-directional coupling*, only the evolving process is influenced by the (static/temporal/multiplex) structure, while the network evolves without being altered by the ongoing phenomena. Good examples are epidemic processes in case patients are not aware of their infection, thus they continue their interaction practise and propagate further the actual disease without intervention [43]. On the other hand, if the system is characterised by *mutual coupling*, node states and interactions influence each other, which leads to adaptive systems with correlations between the phenomena and the network structure. This type of symmetry arguably lays behind the indistinguishability problem of social influence and homophily [262, 26], where one cannot decide whether two connected individuals in a social network are similar because they influenced each other after connection (social influence), or they became connected because they were similar at the outset (homophily). These effects commonly trouble the observation of social spreading phenomena, as we will discuss in the coming Chapter.

Another imaginary dimension may consider the ways of observing a system. While arguably all emerging phenomena in complex systems are the results of some dynamical process, sometimes data do not allow to follow their emergence via *dynamic observations*, but to make only *static observations* on their actual state. While in the former case a possible diachronic analysis may lead to deeper insight about the driving mechanisms (e.g. in case of epidemic spreading), in the latter case observations are limited to the detection of correlations in static informations.

Finally, we may distinguish between different modelling techniques we choose to explain the observed phenomena. One way involves statistical methods, which help us to identify various hidden dependencies. On the other way, we will discuss agent-based models of dynamical processes. Various such processes were proposed to model complex emerging behaviour, e.g. random walks for diffusion, sandpile models for self organised critical systems, or coupled oscillators for synchronisation processes, just to mention a few. Here we will mostly concentrate on spreading phenomena, thus we will limit our discussion to two generic modelling frameworks. On one hand, we will discuss models of *simple contagion processes*, arguably describing epidemic like spreading [43], while on the other hand we will also develop models of *complex contagion processes*, commonly assumed behind social spreading phenomena [78].

The following Chapter is organised via these axes to explore some relation between social networks and ongoing social phenomena. First we will discuss empirical studies based on static or dynamic observations of collective behaviour, and subsequently we will summarise modelling results exploring the critical behaviour and dynamics of simple and complex contagion processes.

4.2 Static observations of collective social phenomena

The un-precedent availability and amount of digital data, recording traces of human behaviour, opened the gate for the quantitative analysis of various social phenomena. Due to these developments it recently became possible (a) to verify earlier social hypothesis, which were taken as granted in the scientific community, but has never been observed quantitatively, (b) to generalise earlier observations made in small-scale experiments usually on a non-representative groups of individuals, and (c) to observe new aspects of human behaviour to build novel hypothesises. Evidently, independently of the method, all such behavioural studies should consider the dynamical aspects of the observed phenomena, however, this is not always the case. Eventually, sometimes it is disregarded for simplicity, or the data records a horizontal view of a social system, involving a large population, but only capturing the state of the system at a given time point.

I have several contributions, which relied on the static observations and analysis of social phenomena. However most of these studies land in the field of socioeconomic networks, which fall somewhat far from a physics description thus I opted to exclude them from this Thesis. They cover several aspects of individual observations and effect of socioeconomic inequalities that characterize any society. First, some are built on directly observed socioeconomic networks, where economic indicators are directly observed for people connected in a social network, like in case of DS2 (for description see Section 1.3.2 where such network is built through the combination of a bank record and a mobile communication dataset. Other studies of mine focus on socioeconomic status inference [245, 244, 14, 137] using remotely sensed satellite images and online data in developed and developing countries. Their common goal is to build socioeconomic maps with high spatial resolution from alternative data sources, this way replacing expensive Census data, especially in developing countries where they are rarely collected. Our methods, as many other in this field, apply deep neural networks to solve this inference problem from extremely large datasets.

Using high resolution socioeconomic maps we coupled them with behavioral information of individuals, such as their social networks [237, 56], linguistic patterns [15], purchase habits [238, 239], or mobility [56, 167]. Commonly, they are investigating socioeconomic stratification patterns indicating segregation in these many aspects of human behavior. On the social network level they appear as assortative contact patterns demonstrating that people of a given socioeconomic class prefer to interact with other from the same or similar classes. This phenomena is reflected by other behavioural aspects too, where shopping patterns or even mobility mixing patterns show the same stratified pattern.

* * *

Finally, note that I have some further contributions related to static analysis of collective social phenomena, which are neither discussed in details here. In one paper [246] we investigate the behavioural differences between mobile phone customers with prepaid and postpaid subscriptions. Our study reveals that (a) postpaid customers are more active in terms of service usage and (b) connections between customers of the same subscription type are much more frequent than those between customers of different subscription types. In another study [269] we address, on a sample of 2.3 million individuals, how Facebook users consumed different information at the edge of political discussion and news during an Italian electoral competition.

4.3 Dynamic observations of social spreading phenomena

In this Section we are going to summarise empirical studies reporting observations about the dynamics of social contagion phenomena. Social contagion evolves over networks of interconnected individuals, where links associated with social ties transferring influence between peers [73]. Several earlier studies aimed to identify the dominant mechanisms at play in such processes [313, 152, 319, 27]. One key element, termed behavioural threshold by Granovetter [152], is defined as *"the number or proportion of others who must make one decision before a given actor does so"*. Following this idea, various network models have been introduced [368, 359, 369, 264, 189] to understand threshold-driven spreading processes, commonly known as *complex contagion* [78]. Although these models are related to a larger set of collective dynamics, they are particularly different from *simple contagion* where the exposure of nodes is driven by independent contagion stimuli and one infected neighbour is always sufficient to expose a susceptible node [48, 43, 166]. During the last ten years several studies contributed to the foundation of complex contagion [368, 30, 141, 314, 37, 324], and in addition online experiments were carried out to provide empirical evidence about the effect of social influence [76, 77]. Beyond the conventional threshold mechanism, the effect of homophily [26, 37, 335] and the role of external media influence [345] were also investigated recently.

Most of my work on dynamical social phenomena are related to the analysis of social contagion processes, with observations in large online social systems like Skype or Twitter. As it follows, we will discuss empirical observations reported in [198, 199, 174] and identify mechanisms, which arguably play central role in the adoption dynamics of online services or information. Subsequently, in Section 4.5, we will incorporate the identified mechanisms into predictive dynamical models of complex contagion processes.

4.3.1 Complex contagion process in spreading of online innovation

The propagation of innovations takes place in a social network [48, 345, 289, 26] and is driven by the entanglement of individuals' decision-making processes [306] as well as by the influence of media and social interactions [313]. Although the effects of network structure on contagion processes have recently been shown to be important [43], knowledge about the social network itself is rather limited since its structure and dynamics usually remain hidden. In this respect the digital age has opened up unprecedented opportunities, as online social networks and Voice over Internet Protocol services record detailed information of the connections and activities of their users. These services partially decode the underlying social structure by acting as proxies for the network of real social ties between individuals, and also provide accurate records of the users' adoption behaviour. In this way the different sources of influence on the decisions of an individual immersed in a perpetually changing environment of social interactions become traceable.

In this project [198] we studied one of today's largest online communication services, the Skype network, with over 300 million monthly connected users. Data covers the history of individuals that have adopted Skype from September 2003 until March 2011 (i.e. 2738 days), including

registration events and contact network evolution for every registered user around the world. For our investigation we selected user accounts with an identified country of registration and considered only their mutually confirmed connections, both within the country and abroad. To receive the best estimation of node degrees in the underlying social network, we integrated the evolving Skype network for the whole available period and count the number of confirmed relationships per node (including international ties). The adoption dynamics of a given country can be directly observed by assigning times of adoption (t_a) and termination (t_t) to all the accounts. These are respectively defined as the dates of registration and last activity (as regards to any of the services) in Skype. Explicitly, we identified any account as terminated if its last activity happened earlier than one year prior to the end of the observation period. In this way we were able to build a complete adoption and termination history of Skype for 2373 days.

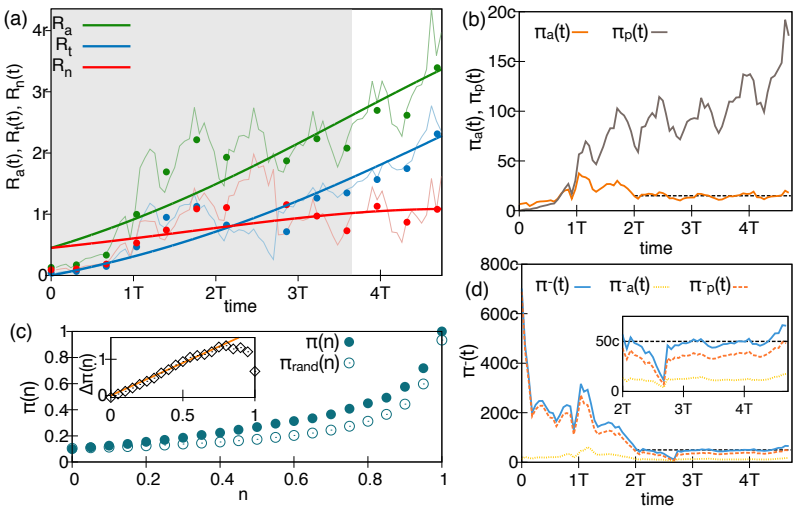


Figure 4.1: Empirical rates and probabilities for Switzerland. (a) Thin curves denote empirical rates of adoption $[R_a(t)]$, termination $[R_t(t)]$, and net adoption $[R_n(t)]$, while symbols are their corresponding binned values. A binned data point in $[2T,3T]$ has been removed due to systematic bias in $R_t(t)$ caused by a major software update during this period. A shaded (white) area indicates the training (predicted) period for the theoretical fit of our model, drawn as thick lines with the same colours as the empirical rates. (b) Probabilities of spontaneous $[\pi_a(t)]$ and peer-pressure $[\pi_p(t)]$ adoption per unit time. (c) Average conditional probability of adoption as a function of the fraction of adopting neighbours n, measured in the original data $[\pi(n)$, solid circles] and in the shuffled data corresponding to the null model $[p_{\text{rand}}(n)$, open circles]. Inset shows the unbiased difference $\Delta\pi(n) = \pi(n) - \pi_{\text{rand}}(n)$ (symbols) and a fitted linear function (continuous line). (d) Probabilities of overall termination $[\pi^-(t)]$, and of spontaneous $[\pi_a^-(t)]$ and peer-pressure $[\pi_p^-(t)]$ termination per unit time. The inset depicts a zoom from time $2T$ onwards. T, r and c are arbitrary linear scaling constants, with time dimensions for T. Black lines in panels (b), (d) are fitted constants. This figure was published in [198].

The adoption dynamics

The spreading of the online service (in this case the adoption of the Skype free service) is determined by competing processes of adoption and termination, measured by the evolution of the corresponding rates $R_a(t)$ and $R_t(t)$ of all users that adopt or terminate the service in a given time window Δt (Fig. 4.1.a). These simple rate functions already disclose interesting features of the adoption dynamics, since their overall growth signals continuously accelerating processes of adoption and termination. Yet the actual time evolution of spreading service is better characterised by the net adoption rate $R_n(t) = R_a(t) - R_t(t)$.

Mechanisms of adoption

Opening a user account constitutes a single event in the decision-making process of an individual that is triggered by either spontaneous decisions, the influence of media or by the social environment [368, 26]. On the other hand, users may terminate their accounts for several reasons including vanishing demand or dissatisfaction, by switching to another product permanently, or by simply abandoning the service with a chance of re-adoption (e.g. due to loss of password or intention for lower monitoring). An analysis of the evolving network structure around a given user can help us to detect some of these scenarios, by observing whether an ego adopted or terminated the product before any of its neighbours did; or else followed the decisions previously made by a fraction of them. In this way we can label the performed action as either spontaneous or driven by peer pressure.

To define the related measures we consider the underlying social network as static, meaning that its evolution requires a much larger temporal scale than the adoption process itself. This static structure is defined as the aggregated social network of Skype at the end of the recorded period, and provides a lower estimate for the total number of friends of each individual. Moreover, we assume that the maximum size of the static social network is the number I of internet users in a given country at the end of the observation period (2011) [10], and thus define $I - N_a(t)$ as the population that has not yet adopted Skype at time t.

Under these assumptions, the probabilities per unit time that a user adopts either spontaneously or due to peer pressure are defined as,

$$\pi_a(t) = \frac{\#ad(t+\Delta t|SF=0)}{I-N_a(t)}, \quad \text{and} \quad \pi_p(t) = \frac{\#ad(t+\Delta t|SF\neq 0)}{I-N_a(t)}, \tag{4.1}$$

where $\#ad(t+\Delta t|SF=0)$ [$\#ad(t+\Delta t|SF\neq 0)$] is the number of users who adopt the service in a time window Δt, under the condition that their number of adopting neighbours at time t is $SF=0$ ($SF\neq 0$). In a similar fashion, the probabilities per unit time that a user terminates the service either spontaneously or due to peer pressure are,

$$\pi_a^-(t) = \frac{\#tr(t+\Delta t, TF=0)}{N_a(t)}, \quad \text{and} \quad \pi_p^-(t) = \frac{\#tr(t+\Delta t, TF\neq 0)}{N_a(t)}, \tag{4.2}$$

where TF stands for the number of neighbours of a user that have terminated usage up to time t (for a discussion on the restrictions of these empirical quantities see [198]).

The data shows that after an initial, transient period, the rate of spontaneous adoption $\pi_a(t)$ (Fig. 4.1b) and the rate of termination $p^-(t) = \pi_a^-(t) + \pi_p^-(t)$ (Fig. 4.1d) become constant apart from small fluctuations. The same holds separately for the rates of spontaneous [$\pi_a^-(t)$] and peer-pressure [$\pi_p^-(t)$] termination. The time invariance of these rates is an obvious assumption for most biological epidemics, which, however, has never been empirically shown before in the case of social contagion phenomena, despite its wide use [364, 73].

When the ego is not the first adopter among neighbours, the rate $\pi_p(t)$ of adoption via peer pressure is not constant but increases with time (Fig. 4.1b). This is arguably due to social influence arising from the user's social circle. An appropriate way to quantify such effects is to measure the conditional probability $\pi(n)$ of adoption provided that a fraction n of the ego's neighbours have adopted the product before as

$$\pi(n) = \frac{\#ad(n)}{N - \sum_{m=0}^{m<n} \#ad(m)}. \tag{4.3}$$

Here the numerator counts the number of users with a fraction n of adopter friends at the time of adoption, while the denominator is the number of people with a larger or equal fraction $m \geq n$, i.e. all individuals who had the chance to adopt Skype while having a fraction n of adopter neighbours.

We observe that the probability $\pi(n)$ is monotonically increasing (Fig. 4.1c), an empirical finding in agreement with the assumptions of several threshold models for epidemic spreading and social dynamics [368, 107, 220, 337]. However, since we cannot see the entire social network (only the part uncovered by the Skype graph), this probability is biased as $n \to 1$. To estimate such bias, we build a reference null model by shuffling the adoption times of all accounts and measuring the corresponding conditional probability $p_{\text{rand}}(n)$ for this system. The shuffling procedure removes the effect of social influence but conserves the adoption rates and keeps the social structure unchanged. In other words, the reference probability is biased in the same way as the original measurement, but is not driven by social influence as all such correlations have been removed by the shuffling. Consequently, the difference $\Delta\pi(n) = \pi(n) - \pi_{\text{rand}}(n)$ quantifies the effect of social influence in the adoption process (inset of Fig. 4.1c): $\Delta\pi(n)$ increases approximately in a linear fashion with the fraction of adopting neighbours. This observation is in agreement with previous studies where a similar scaling of social influence has been recognised through small scale experiments [76], data-driven observations [37], and modelling [107, 106].

Thus based on these observations, for modelling purposes we can summarise that the behaviour of an agent can be characterised by four elementary processes: (a) *Spontaneous adoption*, influenced by individual factors or external media independently of the social network. This is certainly the dominant mechanism for agents with no neighbours at the time of adoption. (b) *Peer-pressure adoption*, an intrinsic social effect implemented here by making use of the observed linear scaling of the probability $\pi(n)$. (c) *Temporary termination*, describing the case in which agents stop usage with a chance of re-adoption. (d) *Permanent termination*, when users abandon the service altogether.

4.3.2 Local cascades induced global contagion

Behavioural cascades are rare but potentially stupendous social spreading phenomena, where collective patterns of exposure emerge as a consequence of small initial perturbations [199]. Some examples are the rapid emergence of political and grass-root movements [149, 61, 117], fast spreading of information [109, 157, 38, 369, 159, 243, 242, 143] or behavioural patterns [128], etc. The characterisation [143, 60, 138, 67, 135] and modelling [368, 324, 141] of such processes have received plenty of attention and provide some basic understanding of the conditions and structure of empirical and synthetic cascades on various types of networks [383, 191, 29]. However, these studies commonly fail in addressing the temporal dynamics of the emerging cascades, which may vary considerably between different cases of social contagion. Moreover, they have not answered why real-world cascades can evolve through various dynamic pathways ranging from slow to rapid patterning, especially in systems where the threshold mechanisms play a role and social phenomena spread globally. Besides the case of rapid cascading mentioned above, an example of the other extreme is the propagation of products in social networks [48], where adoption evolves gradually even if it is driven by threshold mechanisms and may cover a large fraction of the total population [198]. This behaviour characterises the adoption of online services such as Facebook, Twitter, LinkedIn and Skype (Fig.4.2a), since their yearly maximum relative growth of cumulative adoption [376] (which is the maximum of the yearly adoption rate normalised by the final observed adoption number of a given service) is lower than in the case of rapid cascades as suggested e.g. by the Watts threshold model.

To fill this gap in the modelling of social diffusion, we analyse and model real-world examples of social contagion phenomena. We follow the adoption dynamics of the Skype paid service “buy credit” for 89 months since 2004, which evolves over the social network of one of the largest voice over internet providers in the world. Data includes the time of first payment of each user, an individual and conscious action that tracks adoption behaviour. In contrast to other empirical studies where incomplete knowledge about the underlying social network leads to unavoidable

bias [198], in this study, we use the largest connected component of the aggregated free Skype service as the underlying structure, where nodes are Skype users and links confirmed contacts between them. This is a good approximation since it maps all connections in the Skype social network without sampling, and the paid service is only available for individuals already enrolled in the Skype network. Also note that the service adoption process evolves in a considerably faster time-scale than the underpinning social network. This way applying a time-scale separation, and considering the network to be static, may provide a good first approximation here. The underlying structure is an aggregate from September 2003 to November 2011 (i.e. over 99 months) and contains roughly 4.4 billion links and 510 million registered users worldwide [272]. The data is fully anonymised and considers only confirmed connections between users (for more on the data see DS4 in Section 1.3.2).

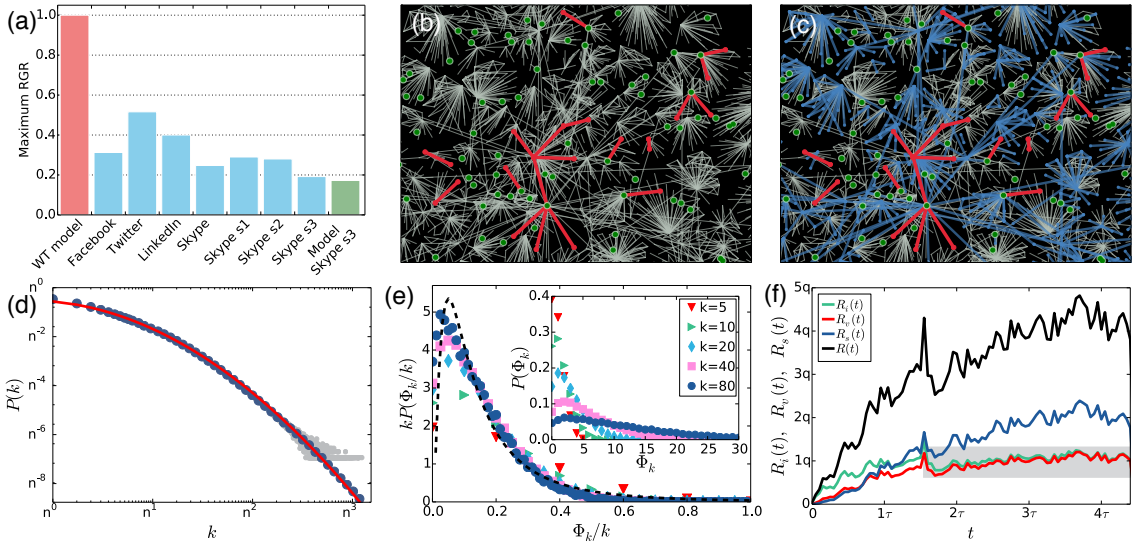


Figure 4.2: Structure and dynamics of online service adoption. (a) Yearly maximum relative growth rate (RGR) of cumulative adoption [199] for several online social-communication services [378], including three Skype paid services (s1 - "subscription", s2 - "voicemail", and s3 - "buy credit"). The red bar corresponds to a rapid cascade of adoption suggested by the Watts threshold (WT) model, while the green bar is the model prediction for Skype s3. (b-c) Snowball sample of the Skype social network (grey links) with nodes and links coloured according to their adoption state: multiple innovators (green nodes), induced small vulnerable trees (red nodes and links), and the triggered connected stable cluster (blue nodes and links). (d) Degree distribution $P(k)$ of the Skype network (grey/blue circles for raw/binned data) on double log-scale with arbitrary base n. $P(k)$ is fitted by a lognormal distribution with parameters $\mu_D = 1.2$ and $\sigma_D = 1.39$, and average $z = 8.56$ (red line). (e) Distribution $P(\Phi_k)$ of integer thresholds Φ_k for several degree groups in Skype s3 (inset). By using $P(\Phi_k, k) = kP(\Phi_k/k)$, these curves collapse to a master curve approximated by a lognormal function (dashed line in main panel) with parameters $\mu_T = -2$ and $\sigma_T = 1$, as constrained by the average threshold $w = 0.19$. (f) Adoption rate of innovators $[R_i(t)]$, vulnerable nodes $[R_v(t)]$, and stable nodes $[R_s(t)]$, as well as net service adoption rate $[R(t)]$. Rates are measured with a 1-month time window, while q and τ are arbitrary constants. The shaded area indicates the regime where innovators adopt approximately with constant rate. This figure was published in [199].

Watts cascade conditions

In his seminal paper about threshold dynamics, Watts [368] classified nodes into three categories based on the necessary social influence they needed for adoption. He assumed that each node has an individual threshold $\varphi \in [0,1]$ drawn from a distribution $P(\varphi)$ with average $\phi = \langle \varphi \rangle$. This threshold determines the minimum fraction of exposed neighbours that triggers adoption and captures the resistance of an individual against engaging in spreading behaviour. He identified

innovator nodes that spontaneously change state to 1, thus starting the process. Such nodes have a trivial threshold $\varphi = 0$. Then there are nodes with threshold $0 < \varphi \leq 1/k$, called *vulnerable*, which need one adopting neighbour before their own adoption. Finally, there are more resilient nodes with threshold $\varphi > 1/k$, denoted as *stable*, referring to individuals in need of strong social influence to follow the actions of their acquaintances. Watts suggested [368] that a small perturbations (like the spontaneous adoption of a single seed node) can trigger global cascading patterns, however, their emergence is subject to the so-called *cascade condition*: the innovator seed has to be linked to a percolating vulnerable cluster, which adopts immediately afterwards and further triggers a global cascade (i.e. a set of adopters larger than a fixed fraction of the finite network). In addition, while models with more sophisticated social influence function can be introduced [231, 107, 198] the original linear-threshold assumption proposed by Watts and Granovetter seems to be sufficient to interpret our observations.

Heterogeneous thresholds and slow adoption dynamics

Degree and threshold heterogeneities are indeed present in the social network of Skype. The degree distribution $P(k)$ is well approximated by a lognormal function

$$P(k) \propto k^{-1} e^{-(\ln k - \mu_D)^2/(2\sigma_D^2)} \quad \text{where} \quad k \geq k_{\min} \tag{4.4}$$

with parameters $\mu_D = 1.2$, $\sigma_D = 1.39$ and $k_{\min} = 1$ (Fig. 4.2d), giving an average degree $z = 8.56$. Moreover, at the time of adoption we can measure the threshold $\varphi = \Phi_k/k$ of a user by counting the number Φ_k of its neighbours who have adopted the service earlier. We then group users by degree and calculate the distribution $P(\Phi_k)$ of the integer threshold Φ_k [138] (Fig. 4.2e). By using the scaling relation $P(\Phi_k, k) = kP(\Phi_k/k)$ all distributions collapse to a master curve well approximated by a lognormal function

$$P(\varphi) \propto \varphi^{-1} e^{-(\ln \varphi - \mu_T)^2/(2\sigma_T^2)}, \tag{4.5}$$

with parameters $\mu_T = -2$ and $\sigma_T = 1$ as constrained by the average threshold $w = 0.19$ (as explained in [199]). These empirical observations, in addition to the broad degree distribution, provide quantitative evidence about the heterogeneous nature of adoption thresholds.

Since we know the complete structure of the online social network, as well as the first time of service usage for all adopters, we can follow the temporal evolution of the adoption dynamics. By counting the number of adopting neighbours of an ego, we identify innovators ($\Phi_k = 0$), and vulnerable ($\Phi_k = 1$) or stable ($\Phi_k > 1$) nodes. The adoption rates for these categories behave rather differently from previous suggestions [368] (Fig. 4.2f). First, there is not only one seed but an increasing fraction of innovators in the system who, after an initial period, adopt approximately at a constant rate. Second, vulnerable nodes adopt approximately with the same rate as innovators suggesting a strong correlation between these types of adoption. This stationary behaviour is rather surprising as environmental effects, like competition or marketing campaigns, potentially influence the adoption dynamics. On the other hand, the overall adoption process accelerates due to the increasing rate of stable adoptions induced by social influence. At the same time a giant adoption cluster grows and percolates through the whole network. Despite of this expansion dynamics and connected structure of the service adoption cluster, the service reaches less than 6% of the total number of active Skype users over a period of 7 years [272]. Therefore we ask whether one can refer to these adoption clusters as cascades. They are not triggered by a small perturbation but induced by several innovators; their evolution is not instantaneous but ranges through several years; and although they involve millions of individuals, they reach only a reduced fraction of the whole network. To answer this question, in Section 4.5 we will incorporate the above mentioned features into a dynamical threshold model [316, 199] with a growing group of innovators and investigate their effects on the evolution of global social adoption. Note that although we cannot follow the

direct pathways of social influence, we performed a null model study to demonstrate at the system level that social influence is present and dominates the contagion process, as compared to effects of homophily (for results see [199]).

4.3.3 Social experiments and behavioral data collection

Large automatically recorded datasets, like the one provided by Skype, serve as important sources for behavioral observations. They allow the representative observations of very large populations as they are collected without intervention as system logs or for billing purposes. Due to these settings, they have minimum observational bias that would alter the behavior of the observed individuals. On the other hand, as their collection is automatized, they provide less flexibility to design problem specific data collection. More adequate solutions, to make observations tailored to answer a specific scientific question, are provided by social experiments or survey studies. In this Section we briefly introduce two projects that we carried out recently in these directions. In one we aim to observe the social and language development of children in a pre-school setting. This is longitudinal and large-scale social experiment is one of the largest of its kind in the field of sociolinguistic. The other data collection effort aim to follow the behavioral changes of people in Hungary during the COVID-19 pandemic. It has been carried out as a continuous online and a sequence of offline survey data collection campaigns, providing the longest dynamical observation of age stratified contact matrices that has been reported so far. These two projects are not closely related to the physical description of social behavior, yet they are related to the overall scope of the thesis on computational human dynamics, thus I decided to introduce them briefly.

DyLNet - Longitudinal data collection to follow social network and language development dynamics at preschool

The aim of the DyLNet project [1] [3] is to observe and characterize the relations between child socialization and oral language learning during the preschool period. This was done by means of an innovative multidisciplinary approach that combines work in the fields of language acquisition, sociolinguistics and network science. The data collection was carried out as a large-scale longitudinal social experiment, where a complete preschool in France was followed, including children from three different grades as well as their teachers and assistants. During the experiment we collected the proximity interactions of about 200 participants (circa 170 preschoolers and 30 adults in charge) in every 5 seconds using autonomous Radio Frequency Identification (RFID) Wireless Proximity Sensors, which were (for a large part) equipped with directional microphones allowing to record continuously the oral interactions of participants too.

More precisely, we collected four different types of datasets during the DyLNet project [96]. The main dataset focuses on the dynamical recording of social and oral interactions as transactional and vocal data. These data were collected autonomously using badges installed on children and school staff at the preschool. Additionally, we tracked the information about the school class level of each child during the observation period together with their teachers and assistants. Meanwhile, ground truth (GT) data were collected with the purpose of understanding how distance and relative orientation between a pair of badges influence the Received Signal Strength Indicator (RSSI) of recorded signals. GT data were also essential for training Machine Learning models to classify signal sequences as social interactions to reconstruct their meaningful interactions as temporal networks [97]. In addition, the main data collection was accompanied with survey campaigns. A first type of survey consisted in asking parents to provide information about the socio-demographic,

[1] The DyLNet project including experimental design, subject recruitment, data collection and processing, data handling, storing and sharing, privacy protection, and all aspects of the involvement of underage children were screened and approved by the ethics committee of INRIA (National Institute for Research in Digital Science and Technology) (favorable opinion, reference 2017-014, IRB00013144) as well as by the Data Protection Officer of the Université Grenoble Alpes (favorable opinion, reference CIL-UGA-2017-0980683).

cultural, educational, and occupational background of the family and the daily out-of-school activities of participating children. In addition, a language survey using vocabulary and syntactic skills assessment methods was performed with all participating children once a year throughout the project to follow their linguistic development. During the observation period, we followed the interactions among preschoolers and school staff for one week each month (among the 10 months of the academic year), for three years of data collections.

MASZK - COVID-19 Hungarian Data Provider Questionnaire

In the early days of March 2020, as the COVID-19 pandemic started to unfold in Hungary, scientists from diverse fields were requested to develop country specific epidemic models. However, soon it turned out that for the realistic parametrization of these models one of the largest challenges was to follow the change of behavior of people in aspects relevant to the transmission of the actual epidemic. Although this effort was supported by a never seen initiative, in which mobile phone providers and health authorities shared their data to help realistic data-driven modelling approaches, one important data was missing from the very beginning: the spatially and demographically detailed mixing patterns of the population's different age groups. Although estimated [304] contact matrices were available for Hungary from earlier periods, the actual challenge was to continuously monitor the changes in contact patterns and to measure the societal responses - like social distancing or self-protection - to the COVID-19 related nationwide regulations. The Hungarian Data Provider Questionnaire ("Magyar Adatszolgáltató Kérdőív" - MASZK) [204, 5] was developed for these purposes[2]. The voluntary and anonymous online survey (designed by scientists and software engineers [6]), is part of a larger project aiming to observe and model the unfolding COVID-19 pandemic in Hungary [315, 298]. Beyond collecting static information about the respondents' demography, domicile, education level, or family structure, the primary goal of the questionnaire was to dynamically monitor the daily changes in the contact pattern of people in order to calculate the age contact matrices in real time. Additionally, dynamic data was collected about the respondents' employment status, working conditions, physical and mental well-being, vaccination acceptance and vaccine preference, and their compliance with other recommended or mandatory self-protection measures during the months of emergency state and beyond. This rolling anonymous online data collection campaign is ongoing up to date (Spring 2022) and reached over 2.3% of the population in Hungary recording over $480,000$ questionnaires from more than $232,000$ individuals, mounting up to the largest data ever collected for this purpose, to our knowledge.

However, as participation was voluntary, just as any data collected in similar ways, the obtained dataset was not representative for the population of Hungary. To estimate the level and dimensions of unrepresentativity, and to have generalizable results, we performed parallel data collection campaigns based on the same questionnaire, but conducted on a smaller representative sample of $1,000-1,500$ people with CATI (computer assisted telephone interviewing) survey methodology in each month from the beginning of the pandemic. This results in 22 representative data collection campaigns up today. Through the combined analysis of the online and offline data, we evaluated the results of the large online survey and identified its most severe non-representative biases. To account for these biases, we developed a pipeline using iterative proportional fitting[53] to weight the non-representative data in order to provide more representative contact matrices. Our method [221] supports the more realistic measurement of age contact matrices of a whole population while keeping the advantages (like cost-efficiency, scalability and detailed dynamics) of the online data collection.

On one hand, the representative contact matrices and the weighted online data were used for weekly updates of the official model predictions of the actual pandemic situation. On the other

[2]The data collection was fully complying with the actual European and Hungarian privacy data regulations and was approved by the Hungarian National Authority for Data Protection and Freedom of Information [7], and also by the Health Science Council Scientific and Research Ethics Committee (resolution number IV/3073- 1 /2021/EKU).

hand, they were exploited for scientific purposes. In a recent study we demonstrated [57] how online data sources can be used as an alternative surveillance system to follow the development of the basic reproduction number during the pandemic. From the analysis of the representative questionnaire [229] we also concluded how the availability and free choice between various vaccine types in Hungary effected the success of the vaccination campaign in different socio-demographic groups. We published a chapter in a White Book on vaccination attitudes in Hungary [248] and finally I co-authored an opinion paper [341] about the socioeconomic divide in the computational modelling of infectious diseases.

* * *

Finally note that I have some further contributions related to the dynamic analysis of collective social phenomena, not discussed in details here. In one study [174] we analysed a Twitter corpus (DS3 in Section 1.3.2) and quantified dynamic similarities between users by considering the evolving set of their common friends and the set of their commonly shared hashtags in order to predict the evolution of mention links among them. We showed that these similarity measures are correlated among connected people and that the combination of contextual and local structural features provides better predictions as compared to cases where they are considered separately.

In a project [103, 104] we collected a specific Twitter dataset to observe the effects of repeated and diversified influence mechanisms my quantifying *Information susceptibility* and *Adoption thresholds* of Twitter users. We track the timeline order of potential influence and adopting hashtags and retweets in a specific Twitter dataset, which contains the posts' dynamics of thousands of seed users and their entire followee networks of over a million of other customers. We show that users adopt retweets easier than hashtags, and we find both metrics to be heterogeneously distributed, correlated, and dependent on the topics and aggregation level of social influence.We show that users adopt retweets easier than hashtags, and find that new influencing neighbours can effectively trigger adoptions. Our results may inform better models of adoption processes leading to a deeper empirical understanding of simple and complex contagion.

In another work [20] we studied multimodal transportation systems, with several coexisting services like bus, tram, and metro of several French municipal areas. Transportation systems can commonly be represented as time-resolved multilayer networks where the different transportation modes connecting the same set of nodes are associated to distinct network layers. In this study, we provided a novel user-based representation of public transportation systems, which combines representations, accounting for the presence of multiple lines and reducing the effect of spatial embeddedness, while considering the total travel time, its variability across the schedule, and taking into account the number of transfers necessary. After the adjustment of earlier techniques to the novel representation framework, using non-negative matrix factorisation, we identified hidden patterns of privileged connections linking places at a given distance with the fastest multimodal transportation connections. We also studied their efficiency as compared to the commuting flow.

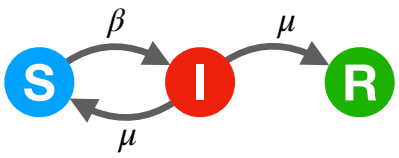
4.4 Modelling simple spreading phenomena

A simplified mathematical description of spreading processes is proposed by compartment models, which commonly assume that at a given time a node can be in one of many mutually exclusive states and can be dynamically transferred between these states following some stochastic or deterministic rules. Based on the compartmental states and the transition rules we distinguish between *simple* [43] and *complex contagion processes* [314, 78] modelling epidemic type of spreading in the former case, while social contagion in the latter. Discussion of my contributions to modelling simple [250, 205, 342, 386] and complex [198, 316, 199, 354] contagion will be the subject of this Section. Our central questions will be to understand the structural effects of temporal

and static networks on spreading processes using synthetic and data-driven models, and to define methods to characterise the underlying network structure sampled by spreading processes.

4.4.1 Simple contagion processes

The spread of infectious diseases are common examples of simple contagion, which can be well described by compartment models of *reaction-diffusion processes* [209, 43]. In such models one takes N nodes, all of which can be in multiple but mutually exclusive states. One commonly used model family of simple spreading processes assumes that nodes can be susceptible (S), infected (I), or recovered/removed (R) and can transfer between these states in consecutive iteration steps via probabilistic rates. If in a given time step a node is susceptible and interacting with an infected other (reaction), in the next iteration step it can become infected (diffusion) with a rate β (described by the spreading scheme $S+I \xrightarrow{\beta} 2I$). Assuming only these two possible states and a single way of transition we obtain the definition of the well known Susceptible-Infected (SI) model. On the other hand, if we allow for an infected node to spontaneously transfer with a given rate μ into other states, we can get more complicated model definitions. In case an infected node can recover ($I \xrightarrow{\mu} R$) and reach a state from which it never get infected again, we arrive to the definition of the Susceptible-Infected-Recovered model. Otherwise, if it can transfer back to state S ($I \xrightarrow{\mu} S$) and became susceptible again for infection, we arrive to the definition of the Susceptible-Infected-Susceptible (SIS) model. Schematic summary of states and transitions scheme of these models are summarised in Fig. 4.3. Although the mathematical definition of such models are very similar, they display rather different critical behaviour when transitioning between phases of vanishing or global contagion [43].



SI	SIS	SIR
$S+I \xrightarrow{\beta} 2I$	$S+I \xrightarrow{\beta} 2I$	$S+I \xrightarrow{\beta} 2I$
	$I \xrightarrow{\mu} S$	$I \xrightarrow{\mu} R$

Figure 4.3: Schematic summary of compartment models of epidemic spreading.

Compartment models of spreading processes were originally defined for unstructured populations assuming *homogeneous mixing*, i.e., where every node interacts with any other nodes in each iteration. In this case, solutions for the dynamics of compartment sizes were proposed by the *dynamical mean-field approach*. However, over the last decades this framework has been successfully extended for structured populations, where a network structure is used to code the possible interactions between nodes. Solution in this case was built on a *degree decomposition method* [293], which introduces classes of statistically equivalent nodes of degree k in the description. This idea lead to the seminal result that degree heterogeneities decrease the critical point of epidemics leading to a vanishing infection threshold in scale-free networks with degree exponent $\gamma \leq 3$.

One common assumption in the majority of related works is to consider a *time-scale separation* between the changes in network structures, τ_G, and the contagion process τ_P (for a discussion on temporal scales see Section 3.1.2). Indeed, spreading processes have been typically considered to take place in either static ($\tau_P \ll \tau_G$) or annealed ($\tau_P \gg \tau_G$) networks, where links are not evolving but present always or they evolve so rapidly that in turn they allow the infection to pass between nodes at any time. While these approximations can be used to study a range of processes such as the spreading of some diseases in contact networks or the propagation of energy in power grids they fail to describe many other phenomena in which the two timescales are comparable [195, 268, 311, 271, 332, 218, 206, 205, 310, 320, 191, 358].

In these cases epidemics may strongly depend on the dynamics of the network, as infections

can be transferred between nodes only at the time of their interactions. Spreading of ideas, memes, information and some type of diseases the diffusion processes can take place in such time-varying networks and their modelling needs to consider the interplay between the two simultaneously unfolding dynamics. These new directions land temporal networks in the focus of epidemic modelling [358, 250] and set a new direction in their investigations called *temporal network epidemiology* [257].

Several other characters of networks have been considered for ever more realistic modelling of epidemic spreading including multi-layer description of social interactions and mobility patterns [151, 387], or meta-population networks [88, 89, 342] to capture human mobility patterns at multiple scales. In the coming Section I am going to summarise some of my contributions in this direction, addressing how to effectively control epidemic spreading on temporal networks [250]. As another example for the usefulness of simple spreading processes, we will also introduce a new centrality measure based on deterministic SI processes to identify important links to impede epidemic outbreaks [386]. Finally we will shortly summarise our other works addressing how epidemic spreading is influenced by effects of contact memory [206] or emerging communities [233] in temporal networks and discuss how the super-linear scaling of contact densities with city sizes influences the outcome of meta-population models of spreading [342].

4.4.2 Controlling contagion processes in time-varying networks

Control strategies of contagion processes is a central question in network epidemiology and in information spreading [86, 294]. The aim here is to identify the minimum set of nodes (or links), which removal would maximally decrease the probability of the emergence of global spreading on one hand, or to reach maximal involvement once initiate spreading from these sites. Spreading processes evolving on networks are influenced by a set of network features. In particular, heterogeneity observed in the distribution of networks' metrics, like the number of connections per node, degree, and the intensity of contacts, or weights have been shown to be critical. These quantities follow distributions characterised by heavy-tails, which imply the absence of characteristic scales and the presence of large fluctuations with respect to the average [363]. Other influencing factors identified are higher-order organisation of connectivity patterns associated to the presence of clusters/communities [278, 126].

The understanding of these properties and their effects on spreading phenomena has spurred the creation of strategies aimed at controlling or promoting diffusion processes. These can be classified in two main categories [366]. On one hand we can devise *global* strategies that rely on the full knowledge of the network structure, while on the other hand we can define *local* strategies, which relax this, often unrealistic assumption. In order to better understand the problem setting, let us imagine that we want to protect a network of computers against the spreading of malwares. The problem is to find a way to immunise a fraction p of nodes to effectively protect the entire network. Each prescription for the selection of this fraction constitutes what we call a strategy. To this end, global strategies use centrality measures such as degree, k-core, betweenness and PageRank to rank the importance of each node [43, 278, 215, 294], while local strategies instead infer the role of nodes by local explorations and samples [86].

Here we investigate the effect of time-varying connectivity patterns on contagion control strategies by considering the specific class of activity driven network models [299]. In particular, we consider the susceptible-infected-susceptible (SIS) model [209] and derive analytically its critical immunisation threshold in case of three different control strategies. We also validate qualitatively the findings obtained in synthetic networks by studying the effect of each strategy in a large-scale mobile telephone call dataset.

Controlling contagion processes in activity-driven networks

A closed formula for the epidemic threshold of a SIS epidemic process unfolding on *any* time-varying network has been derived [303]. In this approach the network is considered as a sequence of adjacency matrices $A_1, A_2, \ldots, A_T$ and has been shown that a disease cannot spread in the system if $\lambda_{\prod_i \mathbf{S_i}} < 1$, where $\mathbf{S_i} = (1-\mu)\mathbf{I} + \alpha \mathbf{A_i}$, and α is the transmission rate per contact. In other words, the disease will die out if the largest eigenvalue of the system-matrix $\mathbf{S} = \prod_i \mathbf{S_i}$ is smaller than one. This result have been recently confirmed with a different approach [358].

In case of activity-driven networks (for definition see Section 3.5), a solution for the SIS model can be obtained by using the homogeneous mean-field theory (discussed for degree decomposition in Section 4.4.1). However, here instead of degrees, we group nodes according to their activity assuming that nodes in the same class are statistically equivalent. At the mean-field level, the spreading process can be described by the number of infected individuals in the class of activity a at time t, i.e., I_a^t. Following Ref. [299], the number of infected individuals of class a at time $t+1$ is given by:

$$I_a^{t+1} = I_a^t - \mu I_a^t + \alpha m (N_a - I_a^t) a \int da' \frac{I_{a'}^t}{N} + \alpha m (N_a - I_a^t) \int da' \frac{I_{a'}^t a'}{N}, \tag{4.6}$$

where N_a is the total number of individuals with activity rate a (which is constant over time). Each term in the Eq. 4.6 has a clear physical interpretation. In fact, the number of infected nodes in the class a at time $t+1$ is given by: the number of infected nodes in this class at time t (first term), minus the number of nodes that recover and going back to the class S_a (second term), plus the number of infected individuals generated when nodes in the class $S_a^t = N_a - I_a^t$ are active and connect with infected nodes in the other activity classes (third term), plus the number of infected nodes generated when nodes in the class S_a^t are linked by active infected nodes in other activity classes. After finding the solution (not shown here [250]), and considering per capita spreading rate $\beta = \alpha \langle k \rangle$ we can write the threshold for the *SIS* process, ξ^{SIS}, as:

$$\frac{\beta}{\mu} \geq \xi^{SIS} \equiv \frac{2\langle a \rangle}{\langle a \rangle + \sqrt{\langle a^2 \rangle}}. \tag{4.7}$$

In other words, the epidemic threshold is a function of the first and second moment of the activity distribution. Due to the co-evolution of the network structure and the spreading processes, the threshold is not depending on time-aggregated metrics such as the degree. It is defined by the interplay between the timescale of the contagion process and the convolution of the network timescales encoded in the moments of the activity distribution.

Let's now study different immunisation strategies. Following Ref. [250] and earlier works on controlling annealed and static networks [294, 86], we will consider three main strategies: random, global and local. In all the cases, we introduce a fraction p of nodes as immunised. To account for this new class of removed nodes, we introduce a new compartment, R, in the classic SIS scheme. Thus, the Eq. (4.6) becomes:

$$I_a^{t+1} = I_a^t - \mu I_a^t + \alpha m (N_a - I_a^t - R_a^t) a \int da' \frac{I_{a'}^t}{N} + \alpha m (N_a - I_a^t - R_a^t) \int da' \frac{I_{a'}^t a'}{N}. \tag{4.8}$$

First, let us consider the random strategy (RS) in which a fraction p of nodes is immunised with a uniform probability (see Fig. 4.4a) [250]. In this case, the system of equations describing the dynamic process in activity-driven networks can be obtained by setting $R_a = pN_a$. The epidemic threshold condition changes as

$$\frac{\beta}{\mu} \geq \xi^{RS} \equiv \frac{1}{1-p} \frac{2\langle a \rangle}{\langle a \rangle + \sqrt{\langle a^2 \rangle}} = \frac{\xi^{SIS}}{1-p}. \tag{4.9}$$

As expected, when a fraction p of nodes is randomly immunised/removed, the epidemic threshold can be written as the threshold with no intervention, ξ^{SIS}, rescaled by the number of nodes still available to the spreading process. Another important quantity is the critical value of immunised/removed nodes, p_c, necessary to halt the contagion process. This quantity is a function of the network's structure and the specific features of the contagion process. The explicit value of p_c can be obtained by inverting Eq. 4.9. In Fig. 4.4a, we plot p_c as a function of β/μ keeping fixed the statistical properties underlying network. The phase space of the diffusion process is divided into two different regions separated by the red solid line that represents p_c as derived by Eq. 4.9. In the region below the curve, the spreading process will take over, $p < p_c$, however, in the region above the curve, the fraction of removed/immunised nodes is enough to completely stop the diffusion process, $p \geq p_c$. To further assess the efficiency of the immunisation strategy in Fig. 4.4d (green triangles), we plot, as a function of the density of removed/immunised nodes p, the ratio I^p_∞/I^0_∞ where I^0_∞ is the asymptotic density of infected nodes when no-intervention is implemented. As shown clearly in the figure, the random strategy allows a reduction in the fraction of infected nodes just for large values of p.

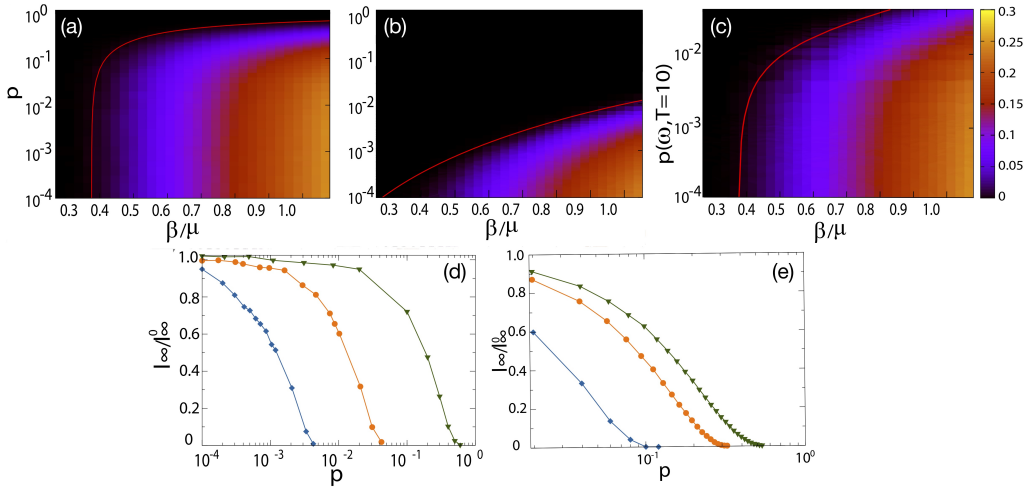


Figure 4.4: Panels a, b, and c show the phase space of an SIS process under random, targeted, and egocentric control strategy, respectively. Considering $N = 10^4$, $m = 3$, $\varepsilon = 10^{-3}$, activity distributed as $F(a) \sim a^{-2.2}$, we plot I_∞ as a function of β/μ and p. Red curves represent the critical value p_c. Panel d shows the comparison of the stationary state of a SIS model with and without control strategy, I^p_∞/I^0_∞, as a function of p when $\beta/\mu = 0.81$. In green triangles, we consider the random strategy, in blue diamonds the targeted strategy, and in orange circles the egocentric strategy. Panel e shows the same results obtained on a real mobile call network. Every simulations was initiated with 1% infected seed, executed 10^2 times for $T = 10^4$ step with $\beta/\mu = 2.5$. Each step was integrated for 6×10^2 seconds, and periodic temporal boundary condition was applied. This figure was prepared by S. Liu and was published in [250].

In networks with heavy-tailed degree distributions, targeting nodes with high degree centrality performs more efficiently than random strategies [43, 85]. Analogously, in activity driven networks, effective strategies shall target high activity nodes. For this reason, we rank nodes in decreasing order of activity and immunising/removing the top ranking pN nodes, and obtain the phase space shown in Fig. 4.4b. This method is equivalent to fix a value a_c so that any node with activity $a \geq a_c$ is immune to the contagion process. Also, for this scheme, it is possible to derive the analytic expression for the epidemic threshold [250]:

$$\frac{\beta}{\mu} \geq \xi^{TS} \equiv \frac{2\langle a \rangle}{\langle a \rangle^c + \sqrt{(1-p)\langle a^2 \rangle^c}}, \tag{4.10}$$

where ξ^{TS} indicates the threshold for the targeted control strategy. In this case it is not possible

to derive explicitly p_c, however, it can be easily evaluated numerically by solving the equation $\xi^{TS} - \beta/\mu = 0$ for different values of β/μ. In Fig. 4.4b, we show p_c (red line) as a function of β/μ. From there it is evident that the immunising/removing a very small fraction of the most active nodes is enough to stop the contagion process, as also confirmed in Fig. 4.4d (blue diamonds). The extreme efficiency of this strategy is due to the crucial role of high activity nodes in the spreading process. Immunising just the top 1% of nodes is enough to halt the disease.

Unfortunately, the network-wide knowledge required to implement targeted control strategies is generally not available [86]. In the case of evolving networks, this issue is even more pronounced as node's characterisation depends on how long it is possible to observe the network dynamics. To solve this problem we propose a local sampling strategy where first we select randomly a w fraction of nodes act as "probes". During an observation time T, we monitor their egocentric network generated by their interactions, which after we select randomly a node in the observed egocentric network of each probe to immunise/remove it. For the sake of comparison with the previous control strategies, we start the epidemic after a p fraction of node have been immunised (for further notes on how to determine p from w see [250]). In case of this egocentric sampling scheme, after some analytical considerations, the epidemic threshold can be expressed as

$$\frac{\beta}{\mu} \geq \xi^{ESS} \equiv \frac{2\langle a \rangle}{\Psi_1^T + \sqrt{\Psi_0^T \Psi_2^T}}, \tag{4.11}$$

where we define $\Psi_n^T = \int da\ a^n (1 - P_a)^T F(a)$. This last integral is a function of the observing time window T, the probability of immunisation/removal of each class, and the activity distribution. We evaluate each Ψ term through numerical integration with results shown in Fig. 4.4c, which make clear that this strategy is much more efficient than the random one, although not as performant as the targeted scheme but with the advantage to rely only on local information. The efficiency of the ES strategy is due to the ability to reach active nodes by a local exploration done observing the systems for few time steps.

Real world time-varying networks add a number of complications to the simplified picture offered by activity driven networks. Indeed, they exhibit correlations among nodes, persistency of links, and burstiness of the activity pattern, as we have discussed it earlier. In order to see whether the above derived mean-field framework provides a good approximation in real world datasets we considered a mobile phone call network (a sample of DS1 in Section 1.3.2), consisting of 93,190 connected phone users of a single city involved in almost five million calls over 120 days. Remarkably, as reported in [250] and shown in Fig.4.4e, we found a very good qualitative agreement between what is observed in real time-varying systems and the analytical results obtained in activity driven networks.

4.4.3 Link transmission centrality in large-scale social networks

Next we discuss another example of simple contagion processes to demonstrate how they can be used to design effective centrality measures in networks [386]. Centrality measures of nodes or links generally rely on local and/or global structural information. Measures using local information, like the node degree or link overlap, are computed efficiently as they only require knowledge about the neighbours of a given node or link. On the other hand, these measures cannot provide information on which nodes or links play global roles in the network structure. On the contrary, centrality measures based on global information about the network structure, like betweenness and closeness centrality [129, 50], Katz centrality [207], k-shell index [58, 215], subgraph centrality [121] and induced centrality measures [122] may better characterise the overall importance of a node or link. Unfortunately, although effective algorithms for approximating these quantities have recently been proposed [65, 119], estimating these measures in large scale networks is still computationally challenging.

While global centrality measures have been very successful in identifying structurally important nodes or links in networks, it has been argued [59] that they do not evidently identify nodes or links with a key role in dynamical processes. Other centrality metrics, which directly use dynamical processes to assign importance, such as PageRank [66], eigenvector centrality [240], or accessibility [347], were found to be more successful in this sense. However, these measures are based on random diffusion processes, and as such they do not fully capture the basic mechanisms behind contagion mediated spreading phenomena. In this Section we define a new link centrality measure, *transmission centrality*, tailored to identify the role of nodes and links in controlling contagion phenomena. The transmission centrality measures the average number of nodes who are reached through each link during the spreading of a stochastic contagion process. This provides a direct measure of the centrality of the link in hindering or facilitating the contagion process.

Transmission centrality

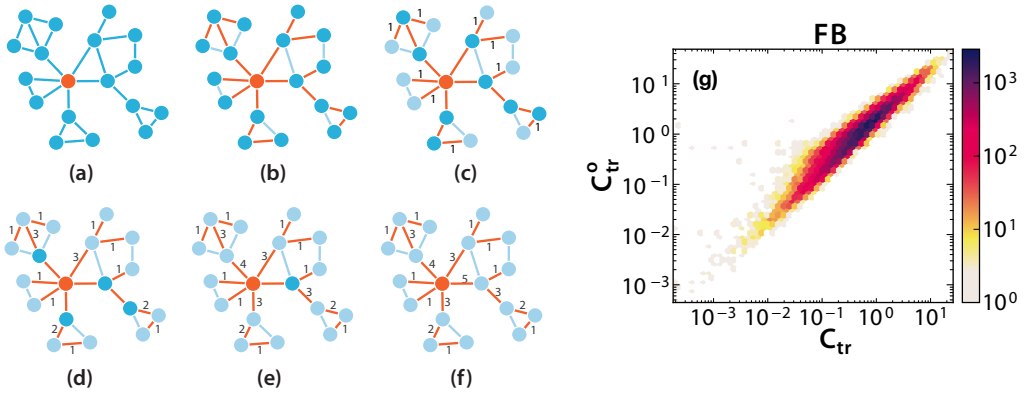


Figure 4.5: Calculation of *transmission centrality* of links. (a) A network with a randomly selected seed node; (b) the branching tree rooted from the initial seed (root and edges in the tree are coloured in red); (c) for each leaf edge in the branching tree increase the counter by 1; (d)-(f) remove leafs and increase the counter of their ascendant by the counter of the removed leafs. (g) Correlation heat-map plot between the C^o_{tr} exact and approximated C_{tr} transmission centrality values of the FB network. Approximate measures were initiated from 5000 seeds and unbiased in $d = 3$ distance. This figure was prepared by Q. Zhang and was published in [386].

Transmission centrality [386] aims to measure for each link in a network its influence in disseminating some globally spreading information. More precisely it measures the number of nodes who received information during a diffusion process through a given link. Its definition intrinsically assumes a diffusion process to unfold on a network structure. In our definition we use the simplest possible information spreading process, the *Susceptible-Infected (SI)* model (for definition see Section 4.4.1 and [43]), however this can be replaced by any other diffusion process. The SI process is controlled by the parameter β, which scales the speed of information/infection spreading. If it is $\beta = 1$ (like in our case by default), it corresponds to the fastest possible information diffusion process determining the shortest diffusion routes between the seed and any other node in the network. In our case, we initiate the SI spreading process from a random seed*s* and simulate the process with a modified breath-first-search algorithm [91, 386]. Using this algorithmic solution one can record the branching tree $G_{BT} = (V_{BT}, E_{BT})$ of the process by keeping track of the direct ascendant of each node from which it received the information. Exploiting the structure of the actual branching tree, *transmission centrality* is formally defined as

$$C_{tr}(u,v) = \begin{cases} \max(|desc(u)|, |desc(v)|), & \text{if } (u,v) \in E_{BT}, \\ 0, & \text{otherwise} \end{cases} \tag{4.12}$$

where $|desc(i)|$ denotes the number of descendant nodes of node i in the branching tree of the actual spreading.

The branching tree G_{BT}, which is a subgraph of G, encodes the diffusion paths that the information takes to reach the vertices of the network. Using its structure we can easily measure the actual C_{tr} value of each link by performing a second step of calculation based on the *river-basin algorithm* [312]. In practice, taking the initial seed s as the root of G_{BT}, and starting from the leafs of the branching tree, we can count the number of descendant nodes of each link, i.e., who received the information via the actual link. The algorithm is summarised in Alg.5, illustrated in Fig.4.5 and works as follows:

Algorithm 5 Transmission centrality

Require: $G = (V,E)$ and $G_{BT} = (V_{BT}, E_{BT})$
Ensure: C_{tr} dictionary of *transmission centrality* values
1: $C_{tr} = dict()$
2: **for** $(u,v) \in G.E$ **do**
3: $\quad C_{tr}((u,v)) = 0$ // set counter to zero for each link
4: **while** $G_{BT}.E_{BT} \neq \emptyset$ **do**
5: $\quad$ **for** $v \in G_{BT}.V_{BT}$ **do**
6: $\quad\quad$ **if** $k_v == 1$ **then**
7: $\quad\quad\quad p = asc(v)$ // parent node of v
8: $\quad\quad\quad gp = asc(p)$ // grandparent node of v
9: $\quad\quad\quad C_{tr}((v,p)) = C_{tr}((v,p)) + 1$
10: $\quad\quad\quad C_{tr}((p,gp)) = C_{tr}((p,gp)) + C_{tr}((v,p))$
11: $\quad\quad\quad G_{BT}.E_{BT} \longleftarrow G_{BT}.E_{BT} - \{(v,p)\}$
12: $\quad\quad\quad G_{BT}.V_{BT} \longleftarrow G_{BT}.V_{BT} - \{v\}$

First we define a dictionary C_{tr}, which associates a counter to each link $(i,j) \in G.E$, that we set to zero initially (lines 1-3 in Alg.5). Then we do the following for every node $v \in G_{BT}.V_{BT}$, which appears with degree $k_v = 1$ in G_{BT}:

(i) Increase by one the counter $C_{tr}((v,p))$ of the (leaf) edge $e_f = (v,p) \in G_{BT}.E_{BT}$, which connects v to its parent node $p = asc_{BT}(v)$ in $G_{BT}.V_{BT}$ (line 9 in Alg.5).

(ii) Increase by $C_{tr}((v,p))$ the counter $C_{tr}((p,gp))$ of its ascendant edge $asc_{BT}(e_f) = (p,gp)$, where $gp = asc(p)$ is the grandparent node of v in $G_{BT}.V_{BT}$ (line 10 in Alg.5).

(iii) Remove v from $G_{BT}.V_{BT}$ and e_f from $G_{BT}.E_{BT}$ (line 11 and 12 in Alg.5). The final transmission transmission centrality value of the actual link $e_f = (v,p)$ is stored in $C_{tr}((v,p))$.

By repeating (a)-(c) for each appearing leaf edge we assign a non-zero value for each link in the branching tree as it is demonstrated in Fig.4.5.c-f.

The transmission centrality of a link can take values between 0 (for links, which are not in the branching tree) and $(N-1)$ (e.g. in the case the seed is propagating information via a single link). Its actual value depends on the choice of the seed node and on the structure of the branching tree determined by the diffusion process. In this way centrality values of the same link may vary from one realisation to another. To eliminate the effects of such fluctuations, the final definition of transmission centrality of links is taken as the average centrality value for each link computed over processes initiated from every node in the network (for a algorithmic definition see [386]). Note that from now on C_{tr} always assigns an average quantity if not stated otherwise.

Heuristic calculation of transmission centrality

One iteration to measure C_{tr} performs with $\mathcal{O}(|E|)$ time complexity, thus in the case when we initiate its calculation from every node $v \in V$, its overall complexity is $\mathcal{O}(|V||E|)$. It is however possible to define a heuristic estimate of transmission centrality at a considerably small computational cost.

As the branching trees of different realisations may largely overlap, a relatively small number of independent realisations, initiated from a reduced set of randomly selected seeds, could provide a good approximation.

Link transmission centrality initiated from a single node provides a locally biased measure as it assigns higher values to links closer to the actual seed. This bias is averaged out if we initiate the spreading process from every node in the network, but in case of a limited number of seeds it has some residual effects. One way to eliminate this residual bias is by assigning zero centrality values to links connecting nodes closer than a distance d to the actual seed. The best value of d depends on the network; however this can be estimated via parameter scanning, as demonstrated in [386].

To illustrate the computation of the heuristic estimate, we use a Facebook network with $20,244$ nodes and $70,132$ edges [386] and compute the average transmission centrality for each link via the exact method by initiating an SI process from each node, and the heuristic method where we initiate processes from 5000 random seeds (i.e. $\sim 25\%$ of all nodes). Further we eliminate biases in distance $d = 3$ around each seed. In Fig. 4.5g we present a heat-map plot about the correlation between the exact (assigned as C^o_{tr} here) and the approximated (assigned as C_{tr}) centrality values of each link. It is evident that there is a strong correlation between these values, quantified by an $r = 0.96$ ($p < 10^{-6}$) Pearson correlation coefficient. Consequently, this unbiased sampling method can provide very close approximations to the exact transmission centrality values, while considerably reducing the computational cost ($\sim 75\%$ in this case).

Due to its advantage to locate central link laying on multiple shortest paths, transmission centrality is very efficient in identifying weak links transmitting information between communities, and which in turn control dynamical processes. In [386] we performed an extensive study to find the best combined tie strength measure, to propose the best link control strategy to impede SIR epidemic spreading. As a result we found that ordering links first by their overlap, and then ordering links having the same overlap values by their transmission centrality provides an efficient method to identify the weakest ties. It is demonstrated there that in some cases the control of only the 30% of the identified weakest links may lead to 90% of reduction of infection size.

Transmission centrality can be generalised in various ways. First, it can be easily defined as a *node centrality metric* by counting for each node the number of their descendant nodes in the branching tree. Moreover it can be extended for *directed and/or weighted networks* by restricting the SI process to respect the direction of links during spreading or by scaling the transmission rate with the normalised weight of links. In addition, for an SI process one can explore central links in the case when the process does not diffuse along the shortest paths. By taking $\beta < 1$, longer spreading paths become plausible allowing the inference of links, which are central in any scenario. Transmission centrality can be easily defined for *temporal networks* [173] as well. Finally, note that transmission centrality as a path based measure is not equivalent but closely related to betweenness centrality. However, while betweenness centrality considers all shortest paths between every pairs of nodes, transmission centrality takes only a single one from the potentially many others. This is especially true when $\beta = 1$, when the SI process is fully deterministic inducing somewhat different but closely matching values with the corresponding betweenness measure, on a considerably lower computational cost.

4.4.4 Switchover phenomenon induced by epidemic seeding on geometric networks

The COVID-19 pandemic highlighted several aspects of real epidemic processes, which were not not in primary focus in earlier studies. Among other, biological, social-behavioral or environmental conditions, the size and location of the seed population where an epidemic is seeded was identified as an important factors that influence the final outcome of the epidemic [265, 89, 25, 214, 95, 133]. Earlier studies suggested, that if the epidemic strikes first at an isolated place with low population density and few local transportation connections, it may become rapidly extinct without causing a

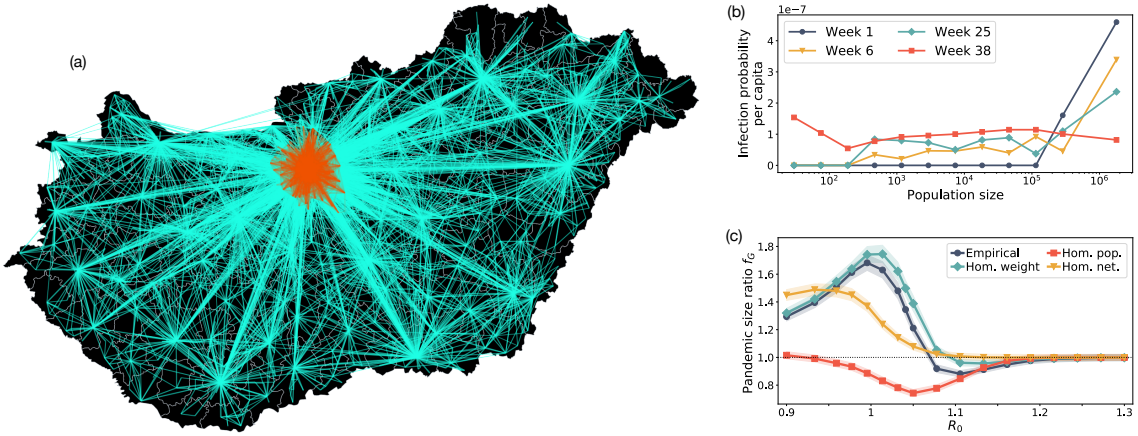


Figure 4.6: Data-driven observations of the switchover phenomenon. (a) Commuting network map of Hungary with settlements larger than 1000 inhabitants and commuting links with more than 25 travelers depicted. Central Hungary (called Center) is highlighted with red. (b) Distribution of per capita infection probabilities in settlements of different sizes at different observation times (in weeks). (c) Simulation results on the real commuting network of Hungary and its three homogenized versions as explained in the main text. Each data point is an average computed from 150 independent simulations, shown with 81% confidence interval. For each of them, initially $s = 97$ settlements are selected according to one of the seeding scenarios (central or uniform). Then, we infect $i_0 = 0.0005$ fraction of the 10^7 agents in the total population, and we distribute these agents in the s settlements uniformly at random, irrespective of the size of the settlements.

major breakout. The dynamics can be entirely different if the epidemic starts from a well connected, more populated place where it can survive and spread to the rest of the population more easily. In a study [285] we challenge this intuition and show that seeding an epidemic from the most tightly connected core of a network does not always lead to a larger epidemic in the long run, in terms of the number of final infected people: If the disease transmits easily, seeding the spreading from nodes selected uniformly at random from the network could reach a larger population.

Empirical observations

The global and local mobility of people are among the most important driving factors behind the spatial spread of most diseases [225, 34, 336]. How people commute locally or travel between cities and countries can be well represented by mobility networks. Concentrating on Hungary, we consider a spatial mobility network (see Fig. 4.6a) describing the average number of daily commuters, who travel to work and school between 1398 settlements with populations larger than 1000 inhabitants according to the 2016 Hungarian micro-census [11]. At the same time, from epidemic data we could follow the daily number of new COVID-19 infection cases in each of these settlements to explore their spatiotemporal distribution in this geometric network.

The first wave of the COVID-19 pandemic started in March 2020 in Hungary. As in many countries, the disease arrived to the country via international air-travel and first landed in larger cities [315, 204, 123, 188] resulting in outbreaks clumped around highly populated areas. This is evident from Fig. 4.6b, where the per-capita infection probability at the beginning of the first wave (week 1) indicates that infection cases were concentrated in cities with the largest populations. Interestingly, this concentration marked the whole 1st wave as even at its peak (week 6 in Fig. 4.6b), most of the infected cases were concentrated in the largest populations, i.e. the capital. On the other hand, at the beginning of the second wave (at the end of August 2020 in Hungary) new infected cases were distributed more homogeneously all around the country. This can be seen from

Fig. 4.6b where the corresponding probability distribution (week 25) is more stretched towards smaller population, as compared to week 1 or 6. This homogenization of infected cases continued during the unfolding of the second wave leading to a fully uniform distribution – corresponding to population densities – at the peak (W38 in Fig. 4.6b). Surprisingly, the first wave that started from the most tightly connected, central, and largest populations led to significantly smaller number of infections as compared to the second wave, that reached an order of magnitude more people, even though it was initiated from more uniformly distributed populations of the network.

Simulation results

To better understand this phenomenon we build a *metapopulation network* [88] using the spatial commuting network of Hungary [11]. In such a network we consider n nodes, which represent populations of individuals (which we also call towns or settlements from now on), connected by weighted edges, encoding the number of people traveling between them. On this network we simulate one of the most basic models for an epidemic spreading [43], an SIR process (for definition see Section 4.4.1). In an SIR model on a metapopulation network, the epidemic evolves in two phases in each iteration. During the *reaction phase*, individuals inside each town mix homogeneously and infect each other following the spreading scheme parametrised with β transmission and μ recovery rates, as explained in Figture 4.3. Subsequently, during the *diffusion phase*, individuals (possibly infected) are selected with probability p_m to move to neighboring nodes in the metapopulation network, this way migrating the epidemic to other towns (for a more formal definition see [285]). Note that the epidemic dynamic is often characterised by the *basic reproduction number*, as defined as the average number of people infected by one ill person in a fully susceptible town ($R_0 = \beta/\mu$).

To capture the observed structural distinction of the central towns in case of the spatial commuting network of Hungary, we identify a *central node set* $\mathscr{C}$, containing the districts of Budapest and its suburbs (red nodes in Fig. 4.6a), which represent about the 30% of the total population of the country [13]. While this definition of $\mathscr{C}$ relies on the specific urban structure of Hungary, we could find more general definitions for $\mathscr{C}$, that are based solely on the network structure. The simplest formal definition would be to take a prescribed number of nodes with the highest degrees. We could also use the core of the network for this purpose, which is obtained by repeatedly deleting all nodes with the lowest degrees as long as only nodes with prescribed degrees remain.

Once selected $\mathscr{C}$, we consider two initial conditions to seed the SIR process in the metapopulation network, starting the spreading from the same number of towns and individuals in both cases. In one case, we choose s ($< |\mathscr{C}| \ll n$) number of towns selected randomly from the $\mathscr{C}$ central set, while in the other case we choose s towns uniformly at random from the whole network. To initiate the spreading, we infect a small i_0 fraction of the total population selected uniformly at random from the chosen s towns, irrespective of their size. This way, for both seeding strategies (centralized or uniform), each seeded town is infected on average with the same number of agents (i_0/s fraction of the total population). To observe the relative effects of the two seeding scenarios, we look at the *experimental pandemic size ratio* $f_G(R_0,s)$, that we define as the ratio of average final infection sizes of epidemic processes seeded from central or uniformly randomly selected towns. Interestingly, as shown in Fig. 4.6c (dark blue line), we find $f_G(R_0,s) > 1$ for small $R_0 \simeq 1$, which means that the epidemics seeded from the $\mathscr{C}$ central set leads to larger outbreaks. However, as we increase R_0, the fraction $f_G(R_0,s)$ falls under 1, thus seeding from uniformly random selected towns over the whole country induces a larger outbreak. This *switchover phenomenon* appears in the slightly super-critical regime, where R_0 is not too large, and where the epidemic never reaches the total population. Instead, due to network effects, it stays clustered around the seeded towns until it dies out. The differences in the infected cluster sizes induced by the two seeding scenarios lead to the observed switchover phenomenon in this regime. On the other hand, if R_0 grows larger, the difference between these seeding scenarios vanishes as the epidemic reaches essentially the whole population in each case.

These observations demonstrate the existence of the switchover phenomenon on a metapopulation model parametrised by the Hungarian commuting network, which is a spatially embedded geometric network (see Fig. 4.6a for Hungary) featuring various structural heterogeneities. To identify which underlying network characteristics are the most important to play a role to induce the observed switchover phenomenon, we use random reference network models [279]. We homogenize the network in different ways to remove certain structural heterogeneities, and compare the outcome of the experimental pandemic size ratio of simulated spreading processes on the randomized structures to our observations on the empirical network (see blue dotted curve in Fig 4.6c). First, to reduce the effects of weight heterogeneities (capturing the variation of number of commuters between pairs of towns), we reset edge weights to the mean weight of all outgoing edges of each node (see green diamond curve in Fig 4.6c). Although this way of homogenization changes somewhat the pandemic size ratio function, it does not have dramatic effects on the observed phenomena. Second, to remove the effects of heterogeneous town sizes and the varying number of commuting individuals from different settlements, we set each town's population to the system average ($\overline{N}$ =6581) and choose the fraction of commuters to be the same (i.e. to $p_m = 0.001$) for each town. Interestingly, this way of homogenization makes the switchover phenomenon even stronger (see red squared curve in Fig 4.6c). Finally, we re-shuffle the ends of network links using the configuration network model [279]. This removes any structural correlations from the network beyond degree heterogeneity, including geometric effects such as long distance connections, the central-periphery structure, structural hierarchy, and locally dense sub-graphs. Due to this shuffling process the switchover phenomenon disappears, or becomes too small to be observed (see yellow triangle curve in Fig 4.6c), indicating that geometric correlations play a central role behind its emergence.

Epidemics and percolation on metapopulation networks

The pandemic size (i.e. the final number of recovered individuals) of a SIR model with deterministic, unit recovery time (e.g. a day) on a (non-meta) network G has a useful connection with the commonly used simple mathematical framework of *bond percolation*. In such a SIR model, every edge of the network G transmits the disease at most once, when one endpoint is infected but the other is still susceptible. Equivalently, one may decide about every edge *in advance*, independently with probability p, whether it will do so. This is called *retaining the edge*, and p is then the retention probability of the model. The retained edges form the percolated random subgraph G^p of G. If a set S of nodes is selected as infected seeds in the network, then the epidemic will spread exactly over the connected components (also called clusters) of G^p that contain at least one node of S.

Metapopulation models are more difficult to treat mathematically, but a fundamental result by [47, 90] connects the behavior of SIR on metapopulation models to bond percolation. Following their arguments, once a large outbreak occurs in a town A, the proportion of infected people within the town *concentrates* around some $r_\infty \in (0,1)$ (called local outbreak ratio). Infected people during the local pandemic carry the infection to a neighboring town B and cause a large outbreak there with a certain – computable – probability:

$$p_{AB} = 1 - \exp\left(-\frac{N p_m w_{AB} r_\infty \left(1 - \frac{1}{R_0}\right)}{\mu}\right), \tag{4.13}$$

where N is the size of each population. Note that the dependence on A and B can be neglected if we assume an unweighted network. Since herd immunity is reached in each town after the first large local epidemic outbreak of size $r_\infty N$, later infections to a town are no longer able to cause macroscopically visible outbreaks. Therefore, after time-rescaling, the towns themselves go through an $S \to I \to R$ progression with unit recovery times and infection probability p. Consequently, the

metapopulation model can be approximated by a simple SIR model on the network of towns, and in turn with a bond percolation process with retention probability p.

The connection between metapopulation models and bond percolation allows us to understand the switchover phenomenon of the pandemic size ratio using a theoretical analysis of percolation cluster sizes, which show a remarkable *phase transition* in the edge retention probability p around a critical point p_c.

Beyond percolation cluster sizes, we also need to understand how the different seedings (central or uniform) interact with the clusters to explain the switchover phenomenon of the pandemic size ratio. In [285] we define two approaches to establish theoretically the observed phenomenon. In the *weak switchover phenonema* we require that there exists a seed count $s \leq n$ and link-retention probabilities $0 < p_1, p_2 < 1$ with

$$f_G(p_1, s) > 1 + c, \quad \text{and} \quad f_G(p_2, s) < 1 - c, \tag{4.14}$$

for some constant c that might depend on the network size. Meanwhile, in the *strong switchover phenomenon*, we require that the constant c does not depend on the network size n.

For simplicity, instead of assuming a geometric network structure, we borrow concepts from a simpler conventional network model, called Stochastic Block Model (SBM), to observe the strong switchover phenomenon. The SBM is able to mimic the central and rural areas of a town network, since it contains a 'hidden geometry'. Using this structure we were able to proof that in the Stochastic Block Model with appropriately scaled parameters and $s_n = \Theta(n)$ the strong switchover phenomenon happens. In addition, we could proof the existence of the weak switchover in case of the ageometric configuration model, already observed experimentally in Fig. 4.6c.

For geometric networks with various node degree distributions, *critical exponents* have been already proposed earlier [281, 84], with some of them proven rigorously for several types of network structures. Based on these results, we also demonstrated that, after appropriate scaling, the pandemic size ratio f_G of the configuration model converges to a two-dimensional limit function, which can be precisely determined [285].

Without detailed proofs, the reasons behind the emergence of the switchover phenomenon can be accounted to the overlapping percolation clusters that seeds could induce initiated from given populations. If the system is in the slightly super-critical regime, multiple seeds which are located in the largest and best connected populations will induce outbreaks that may reach largely overlapping sets of individuals. Thus these outbreaks will disturb each other due to the large overlap of their percolation clusters, which in turn will disadvantage the overall epidemic process. On the contrary, when the same epidemic starts from seeds homogeneously distributed in any population (likely from the periphery of the network), the induced percolation clusters will overlap to a smaller extent. At the same time they will reach the largest populations in a few iterations due to the hierarchical structure exhibiting a small-world property. This way in addition to the less overlapping potentially infected population they will have the same advantage that central seeding strategy could offer. Consequently, they could induce larger infected populations under the same epidemic conditions, when the spreading process is slightly above its critical point.

In summary, in this study we were interested in the long-term behavior of spreading processes and showed that the relative danger of infecting a larger population when starting the process from the core or uniformly at random in a network has a non-monotonous dependency on R_0. We explored an entirely new switchover phenomenon and demonstrated them on real and synthetic networks via numerical simulations. We provided a rigorous proof for the existence of this phenomenon on a large set of random graphs, while we are confident that our theory can be extended for a more general set of graphs, which resembles certain structural constraints. Importantly, we identified the spatial geometry of the underlying structure as an important amplifying factor of the switchover phenomenon. Beyond scientific merit, our results may contribute to the better designs of epidemic

forecasts and intervention strategies in a country during an ongoing pandemic. We highlight the importance to follow not only the rate but also the spatial distribution of new infection cases of a spreading disease or its variants during the early phase of an epidemic.

* * *

In addition to the above studies me and colleagues published several other contributions addressing simple contagion processes on networks. In one line of research [342] we studied the dynamics of reaction-diffusion processes (in particular the spreading of SIR processes) on heterogeneous meta-population networks where interaction rates scale with subpopulation sizes. We presented a new empirical evidence, based on the analysis of the interactions of 13 million users on Twitter, about the scaling of human interactions with population size. We found that they scale super-linearly with an exponent γ ranging between 1.11 and 1.21, as observed in recent studies based on mobile phone data. We integrated these observations into a reaction-diffusion meta-population framework and provided an explicit analytical expression for the global invasion threshold. Interestingly, we found that the super-linear scaling of human contacts facilitate the spreading dynamics. This behaviour is enhanced by increasing heterogeneities in the mobility flows coupling the subpopulations. Our results show that the scaling properties of human interactions can significantly affect dynamical processes mediated by human contacts such as the spread of diseases, ideas and behavioural patterns.

In some other works, we studied simple contagion processes evolving on *temporal networks*. In one set of works [195, 218] we focused on data-driven simulations of information diffusion on random reference models of temporal networks (as already discussed in Section 3.3) to identify relevant structural and temporal correlations which influence the dynamics of the process. In this case we used deterministic SI spreading (with $\beta = 1$) as a stereotypic model for information diffusion and showed that bursty temporal heterogeneities in contact dynamics and Granovetterian weight-topology correlations are mostly responsible for the observed slow spreading dynamics. In another line of works, we studied the effects of contact memory and communities on rumour and information spreading processes evolving on synthetic temporal networks simulated by the *activity-driven network* model. In one case, as we discussed in Section 3.5.2 and in [206], we incorporated a simple statistical law that characterises memory in the temporal evolution of users' egocentric networks. We encoded this mechanism in a reinforcement process defining a time-varying network model that exhibits the emergence of strong and weak ties. On this network we studied the effects of time-varying and heterogeneous interactions on the classic rumour spreading model in both synthetic, and real-world networks. The model used here was the Delay-Kendall model [99], which is very similar to an SIR spreading, except that transition to a removed (stifler) state is not spontaneous but due to some interaction with other infected or removed nodes. We observe that strong ties severely inhibit information diffusion by confining the spreading process among agents with recurrent communication patterns. This provided the counterintuitive evidence that strong ties may have a negative role in the spreading of information across networks.

4.5 Modelling complex spreading phenomena

There are remarkable analogies between the social contagion of information, behavioural patterns or innovation and some physical or epidemic spreading processes, where global phenomena emerge through the diffusion of microscopic states [43, 115, 145, 178]. All evolve in networks with nodes characterised by relevant state variables, and links that represent direct interactions between nodes. In biological systems epidemics are driven by binary interactions that lead to the emergence of *simple contagion* phenomena [43], as we discussed before in Section 4.4.1. On the other hand, social diffusion processes are usually characterised by *complex contagion* mechanisms,

where node states are determined by comparing individual thresholds with all neighbour states [115, 198, 368, 372, 78]. This property, capturing the effect of peer pressure and commonly assumed in social spreading phenomena [154, 76], has consequences on the dynamics and the final outcome of the social contagion process. Moreover, the theoretical approach to these systems has much in common [43, 368, 141], which greatly helps us to understand their behaviour. However, as we have already denoted in Section 4.3, real world observations of social spreading processes suggest significantly different picture as drawn from their modelling. In the following I am going to present a series of studies we published to introduce a framework based on the original model design of Watts [368] and the formalism known as the *approximate mean-field equations* introduced by Gleeson [139, 140]. Our aim with these contributions [198, 316, 199, 354, 355] was to extend and re-define a threshold driven modelling framework, to better capture the dynamics of social contagion and its dependences on structural heterogeneities.

4.5.1 Complex contagion processes

Models employing threshold mechanisms mostly focus on cascading phenomena where, under some circumstances, a macroscopic fraction of nodes in the network is converted rapidly due to microscopic perturbations. This approach is motivated by earlier social theories [154, 319] and has been implemented by Watts in an elegant model of cascading behaviour [368]. Watts showed that a global cascade (occupying a macroscopic fraction of the network and induced by local perturbations) can occur due to the interplay between network structure and individual thresholds (as briefly discussed in Section 4.3.2). He further identified the phase with a non-zero probability of global cascades in the space (ϕ, z) of the average threshold ϕ of nodes and the average degree z of the network.

Watts' threshold model [368] is defined on networks where nodes are associated to a state and can observe the state of their network neighbours at any time during the process. All nodes are initially in a susceptible state 0, except for a single adopter seed in state 1. The process evolves as each node with degree k changes its state from 0 to 1 if a fraction φ of its neighbours have adopted before (as demonstrated in Fig. 4.7). Since nodes cannot change their state after exposure, the system evolves towards a state where no further adoptions are possible. The emergence of a global cascade depends on the degree distribution $P(k)$ of the network, the distribution $P(\varphi)$ of individual thresholds, and the initial seed. As we have already discussed in Section 4.3.2, the condition for a global cascade is the existence of a percolating component of *vulnerable* nodes with thresholds $0 < \varphi \leq 1/k$ (who need one adopting neighbour before exposure) connected to the innovator seed with threshold $\varphi = 0$. This percolating vulnerable tree is quickly converted after adoption of the seed and may trigger further adoption of *stable* nodes with thresholds $\varphi > 1/k$ (who need more than one adopting neighbour to adopt). Assuming an Erdős-Rényi (ER) random network [120] and a single adopter seed, there is a phase boundary in (ϕ, z)-space encompassing a regime where global cascades occur. The properties of this cascading regime have been investigated for the case of heterogeneous thresholds, different network topologies [368, 141], and variable seed size [138, 324].

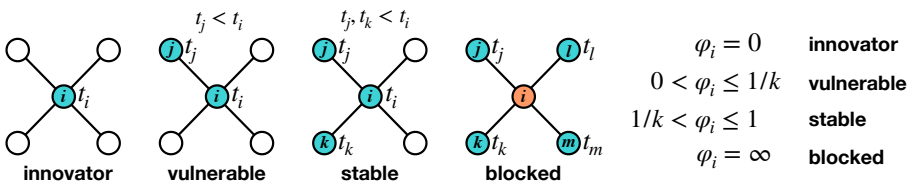


Figure 4.7: Schematic summary of (non-)adopter types and threshold driven transitions in complex contagion processes. Colours assign states of nodes as white-susceptible, blue-adopter and orange-blocked.

While the relevance of the Watts' model is indisputable [368, 141, 369, 140, 139, 138, 383, 324], its limitations become clear from real social spreading data. The Watts model focuses on the (instantaneous) emergence of global cascades triggered by single local perturbations, while there are empirical examples where threshold mechanisms do play a role yet global adoption phenomena emerge through other scenarios (for examples see Fig. 4.2a). Contrary, the Watts criterion for macroscopic adoption is purely deterministic, coded in the network structure, threshold distribution and perturbation site – it does not concern time, which is clearly a feature of empirical stochastic processes of adoption spreading. To resolve this shortcoming, we extend the conventional Watts model with empirically motivated mechanisms (some already reported in Section 4.3.2) to see whether they are eligible to explain the observed dynamical scenarios of social contagion phenomena.

Empirical studies support the intuition that some individuals in the social network may refuse to adopt technological innovations for various reasons – due to another favourite product, aversion towards a firm, or some criticism on principle [199]. Such individuals will never be exposed, irrespective of the state of their neighbours [385], thus their threshold can be regarded as $\varphi = \infty$ (as demonstrated in Fig.4.7). To consider this behavioural pattern we introduce a third state of nodes who are blocked (immunised) to participate in the spreading process. We block the adoption of a fraction r of randomly selected nodes in the network, who in turn do count when their neighbours consider the decision to adopt, and thus will make it harder for neighbours to fulfil the threshold criterion. While blocked nodes hinder the spreading process, there are reasons other than social influence that could motivate individuals to adopt a social pattern, like external influence from mass media. This *spontaneous* adoption has been studied theoretically by introducing a given density of adopters at the outset of the Watts model [324]. However, spontaneous adopters may get active at any time during a real social contagion. Thus we include a stochastic dynamics where a susceptible node may become adopter with rate p at any time, irrespective of the status of its neighbours. This assumption is based on our observations reported in Section 4.3.2 and [199], where after an initial period, innovators adopted a service with an approximately constant rate.

Considering both extensions, we have a threshold-driven dynamics with three node states: blocked, susceptible and adopter (Fig.4.7). At the outset, all nodes are susceptible except for a fraction r that remains blocked. At each time step of the simulation, a randomly selected susceptible node i adopts spontaneously with probability p, otherwise it adopts if at least a fraction φ of its neighbours has already adopted. If $r = 0$ and $p > 0$ all nodes will eventually adopt (Fig.4.8c), following a kinetics reminiscent of the approach to a unique ground state in a physics system. On the other hand, if we introduce quenched randomness and stochastic perturbations ($r, p > 0$), our model allows various temporal regimes and a transition from rapid to slow spreading dynamics.

4.5.2 Dynamical threshold model with immune nodes

Our threshold model can be studied analytically by extending the framework of *approximate master equations* (AMEs) for monotone binary-state dynamics developed by Gleeson [140, 139, 138], where the transition rate between susceptible and adoption states only depends on the number m of neighbours that have already adopted. We describe a node by the property vector $\mathbf{k} = (k, c)$, where $k = k_0, k_1, \ldots k_{M-1}$ is its degree and $c = 0, 1, \ldots, M$ its type, i.e. $c = 0$ is the type of the fraction r of immune nodes, while $c \neq 0$ is the type of all non-immune nodes that have threshold φ_c. In this way, $P(\varphi)$ is substituted by the discrete distribution of types $P(c)$ (for $c > 0$). The integer M is the maximum number of degrees (or non-zero types) considered in the AME framework, which can be increased to improve the accuracy of the analytical approximation at the expense of speed in its numerical computation.

We characterise the static social network by the extended distribution $P(\mathbf{k})$, where $P(\mathbf{k}) = rP(k)$ for $c = 0$ and $P(\mathbf{k}) = (1-r)P(k)P(c)$ for $c > 0$. Non-immune and susceptible nodes with property

vector $\mathbf{k}$ adopt spontaneously with a constant rate p, otherwise they adopt only if a fraction φ_c of their k neighbours has adopted before. These rules are condensed into the probability $F_{\mathbf{k},m}dt$ that a node will adopt within a small time interval dt, given that m of its neighbours are already adopters,

$$F_{\mathbf{k},m} = \begin{cases} p_r & \text{if} \quad m < k\varphi_c \\ 1 & \text{if} \quad m \geq k\varphi_c \end{cases}, \quad \forall m \quad \text{and} \quad k, c \neq 0, \tag{4.15}$$

with $F_{(k,0),m} = 0 \; \forall k, m$ and $F_{(0,c),0} = p_r \; \forall c \neq 0$ (for immune and isolated nodes, respectively). The rescaled rate $p_r = p/(1-r)$ (with $p_r = 1$ for $p > 1-r$) is necessary if we wish to obtain a rate p of innovators for early times of the dynamics, regardless of the value of r.

The dynamics of adoption is well described by an AME for the fraction $s_{\mathbf{k},m}(t)$ of $\mathbf{k}$-nodes that are susceptible at time t and have $m = 0, \ldots, k$ adopting neighbours [302, 140, 139],

$$\dot{s}_{\mathbf{k},m} = -F_{\mathbf{k},m} s_{\mathbf{k},m} - \beta_s (k-m) s_{\mathbf{k},m} + \beta_s (k-m+1) s_{\mathbf{k},m-1}, \tag{4.16}$$

where

$$\beta_s(t) = \frac{\sum_{\mathbf{k}} P(\mathbf{k}) \sum_m (k-m) F_{\mathbf{k},m} s_{\mathbf{k},m}(t)}{\sum_{\mathbf{k}} P(\mathbf{k}) \sum_m (k-m) s_{\mathbf{k},m}(t)}, \tag{4.17}$$

and the sum is over all the degrees and types, i.e. $\sum_{\mathbf{k}} \bullet = \sum_k \sum_c \bullet$. To reduce the dimensionality of Eq. (4.16), we consider the ansatz

$$s_{\mathbf{k},m}(t) = B_{k,m}[\nu(t)] e^{-p_r t} \quad \text{for} \quad m < k\varphi_c \quad \text{and} \quad c \neq 0, \tag{4.18}$$

with $\nu(t)$ the probability that a randomly-chosen neighbour of a susceptible node is an adopter.

Introducing the ansatz of Eq. (4.18) into the AME system of Eq. (4.16) leads to the condition $\dot{\nu} = \beta_s (1-\nu)$. With some algebra, the AMEs for our dynamical threshold model are reduced to the pair of ordinary differential equations

$$\dot{\rho} = h(\nu, t) - \rho, \tag{4.19a}$$

$$\dot{\nu} = g(\nu, t) - \nu, \tag{4.19b}$$

where $\rho(t) = 1 - \sum_{\mathbf{k}} P(\mathbf{k}) \sum_m s_{\mathbf{k},m}(t)$ is the fraction of adopters in the network, and the initial conditions are $\rho(0) = \nu(0) = 0$. Here,

$$h = (1-r)\Big[f_t + (1-f_t) \sum_{\mathbf{k}|c\neq 0} P(k) P(c) \sum_{m \geq k\varphi_c} B_{k,m}(\nu) \Big], \tag{4.20}$$

and

$$g = (1-r)\Big[f_t + (1-f_t) \sum_{\mathbf{k}|c\neq 0} \frac{k}{z} P(k) P(c) \sum_{m \geq k\varphi_c} B_{k-1,m}(\nu) \Big], \tag{4.21}$$

where $f_t = 1 - (1-p_r) e^{-p_r t}$, and $B_{k,m}(\nu) = \binom{k}{m} \nu^m (1-\nu)^{k-m}$ is the binomial distribution. The fraction of adopters ρ is then obtained by solving Eq. (4.24) numerically. Since the susceptible nodes adopt spontaneously with rate p, the fraction of innovators $\rho_0(t)$ in the network is given by

$$\rho_0(t) = p_r \int_0^t [1 - r - \rho(\tau)] d\tau. \tag{4.22}$$

We may also implement the model numerically via a Monte Carlo simulation in a network of size N, with a log-normal degree distribution and a log-normal threshold distribution as observed empirically in the case of Skype (see Section 4.3.2). Hence, we can explore the behaviour of the

fractions of adopters and innovators in the network, ρ and ρ_0, as a function of z, ϕ, p and r, both in the numerical simulation and in the theoretical approximation given by Eqs. (4.24) and (4.22). For $p > 0$ some nodes adopt spontaneously as time passes by, leading to a frozen state characterised by the final fraction of adopters $\rho(\infty) = 1 - r$. However, the time needed to reach such a state depends heavily on the distribution of degrees and thresholds, as indicated by a region of large adoption ($\rho \approx 1 - r$) that grows in (ϕ, z)-space with time (contour lines in Fig. 4.8a). If we fix the time in the dynamics and vary the fraction of immune nodes instead, this region shrinks as r increases (contour lines in Fig. 4.8b). In other words, the set of networks (defined by their average degree and threshold) that allow the spread of adoption is larger at later times in the dynamics, or when the fraction of immune nodes is small. When both t and r are fixed, the normalised fraction of adopters $\rho/(1-r)$ gradually decreases for less connected networks with larger thresholds (surface plot in Fig. 4.8a and b).

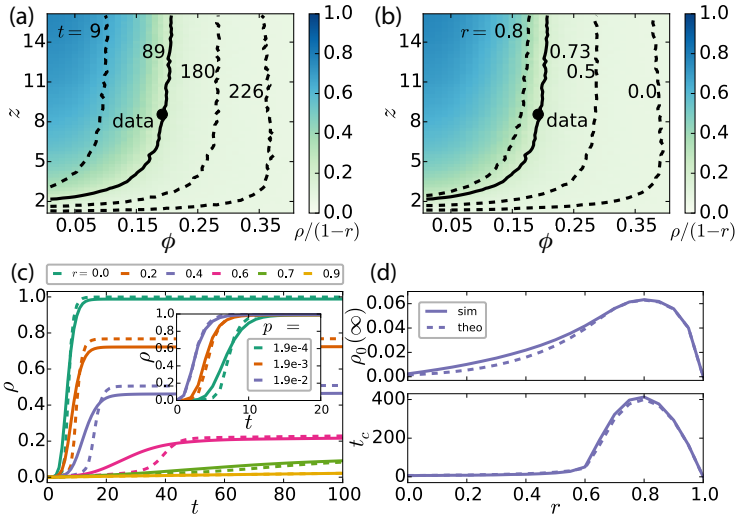


Figure 4.8: (a-b) Surface plot of the normalised fraction of adopters $\rho/(1-r)$ in (ϕ, z)-space, for $r = 0.73$ and $t = 89$. Contour lines signal the parameter values for which 20% of non-immune nodes have adopted, for fixed r and varying time (a), and for fixed time and varying r (b). The continuous contour line and dot indicate parameter values of the last observation of Skype s3. A regime of maximal adoption ($\rho \approx 1 - r$) grows as time goes by, and shrinks for larger r. (c) Time series of the fraction of adopters ρ for fixed $p = 0.00019$ and varying r (main), and for fixed $r = 0$ and varying p (inset). These curves are well approximated by the solution of Eq. (4.24) for $k_0 = 3$, $k_{M-1} = 150$ and $M = 25$ (dashed lines). The dynamics is clearly faster for larger p values. As r increases, the system enters a regime where the dynamics is slowed down and adopters are mostly innovators. (d) Final fraction of innovators $\rho_{0,\infty}$ and the time t_c when 50% of non-immune nodes have adopted as a function of r, both simulated and theoretical. The crossover to a regime of slow adoption is characterised by a maximal fraction of innovators and time t_c. Unless otherwise stated, $p = 0.00019$ and we use $N = 10^4$, $\mu_D = 1.09$, $\sigma_D = 1.39$, $k_{min} = 1$, $\mu_T = -2$, and $\sigma_T = 1$ to obtain $z = 8.56$ and $\phi = 0.19$ as in Skype s3. The difference in μ_D between data and model is due to finite-size effects. Numerical results are averaged over 10^2 (a-b) and 10^3 (c-d) realisations. This figure was prepared by G. Iñiguez and was published in [199].

Both numerical simulations and analytical approximations show how the dynamics of spreading changes by introducing immune individuals in the social network. For $r \approx 0$, the adoption cascade appears sooner for larger p, since this parameter regulates how quickly we reach the critical fraction of innovators necessary to trigger a cascade of fast adoption throughout all susceptible nodes (Fig. 4.8c, inset). Yet as we increase r above a critical value r_c (and thus introduce random quenching), the system enters a regime where rapid cascades disappear and adoption is slowed down, since stable nodes have more immune neighbours and it is difficult to fulfil their threshold

condition. The crossover between these fast and slow regimes is gradual, as seen in the shape of ρ for increasing r (Fig. 4.8c, main panel). We may identify r_c in various ways: by the maximum in both the final fraction of innovators $\rho_{0,\infty} = \rho_0(\infty)$ and the critical time t_c when $\rho = (1-r)/2$ (Fig. 4.8d), or as the r value where the inflection point in ρ disappears. These measures indicate $r_c \approx 0.8$ for parameter values calibrated with Skype data. All global properties of the dynamics (like the functional dependence of ρ and ρ_0) are very well approximated by the solution of Eqs. (4.24) and (4.22) (dashed lines in Fig. 4.8c and d). Indeed, the AME framework is able to capture the shape of the ρ time series, the crossover between regimes of fast and slow adoption, as well as the maximum in $\rho_{0,\infty}$ and t_c.

Thus with our new model extensions we have shown, that outside of the cascading regime of the Watts model, there is possibility of global contagion mediated by spontaneous adopters. However, the speed of spreading depends strongly on the density of blocked or immune nodes. For a small fraction r of blocked nodes, few spontaneous adopters enable the formation of large clusters by initiating cascades. For large r, spreading slows down as it is dominated by spontaneous adopters as suggested by the empirical observations in Section 4.3.2. This way, our intrinsically dynamic model shows a novel percolation transition of induced clusters and is able to describe various scenarios of real social contagion as well as the crossover between them.

Model verification via data-driven simulations

After we defined our dynamical threshold model, let's return back to our empirical observations on the spreading of Skype services reported in Section 4.3.2. As demonstrated above, our model provides insight on the role of innovators and immune nodes in controlling the speed of the adoption process. However, in empirical datasets information about the fraction of non-adopters is usually not available, which makes it difficult to predict the future dynamics of service adoption. Here we use our modelling framework to perform data-driven simulations for two reasons: (a) to estimate the fraction r of immune nodes in the real system; and (b) to validate our modelling as compared to real data.

To set up our data-driven simulations we use the Skype data to directly determine all model parameters, apart from the fraction r of immune nodes. As we already discussed in Section 4.3.2, the best approximation of the degree distribution of the real network is a log-normal function (Eq. 4.4) with parameters $\mu_D = 1.2$, $\sigma_D = 1.39$, minimum degree $k_{min} = 1$ and average degree $z = 8.56$. To account for finite-size effects in the model results for low N, we decrease μ_D slightly to obtain the same value of z as in the real network. We also observe in Fig. 4.2e that the threshold distribution of each degree group collapses into a master curve after normalisation by using the scaling relation $P(\Phi_k, k) = kP(\Phi_k/k)$. This master curve can be well-approximated by the log-normal distribution shown in Eq. 4.5, with parameters $\mu_T = -2$ and $\sigma_T = 1$ as determined by the empirical average threshold $\phi = 0.19$ and standard deviation 0.233. We estimate a rate of innovators $p = 0.00019$ by fitting a constant function to $R_i(t)$ for $t > 2\tau$ (Fig. 4.2f). The fit to p also matches the time-scale of a Monte Carlo iteration in the model to 1 month. To model the observed dynamics and explore the effect of immune nodes, we use a configuration-model network [278] with log-normal degree and threshold distributions and p as the constant rate of innovators, all determined from the empirical data. Model results in Fig. 4.9 are averaged over 100 networks of size $N = 10^5$ (10^6) after $T = 89$ iterations, matching the length of the observation period in Skype.

As a function of r, the underlying and adoption networks pass through three percolation-type phase transitions. First, the appearance of immune nodes (for increasing r) can be considered as a removal process of nodes available for adoption from the underlying network structure. After the appearance of a critical fraction of immune nodes, r_c^{net}, the effective network structure available for adoption will be fragmented and will consist of small components only, limiting the size of the largest adoption cluster possible. Second, r also controls the size of the emergent adoption cascades evolving on top of the network structure. While for small r the adoption network is connected into

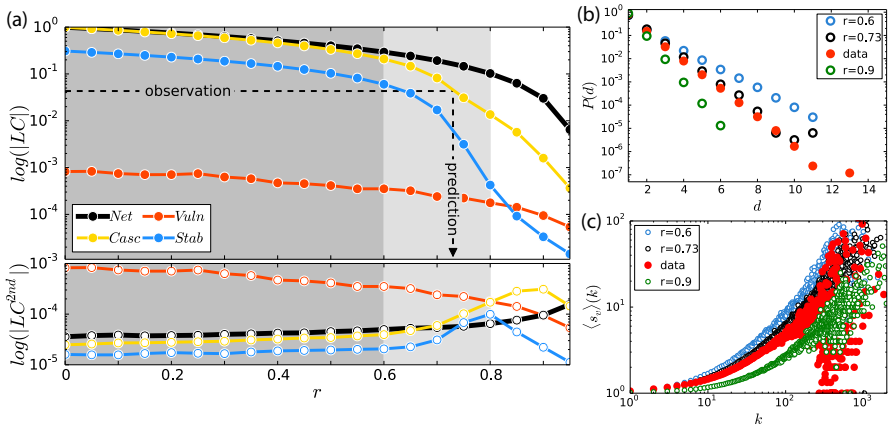


Figure 4.9: (a) Average size of the largest (LC) and 2nd largest (LC^{2nd}) components of the model network ('Net'), adoption network ('Casc'), stable network ('Stab'), and induced vulnerable trees ('Vuln') as a function of r. Dashed lines show the observed relative size of the real LC of the adopter network in 2011 and the predicted r value. (b) Distribution $P(d)$ of depths of induced vulnerable trees in both data and model for several r values, showing a good fit with the data for $r = 0.73$. The difference in the tail is due to finite-size effects. (c) Correlation $\langle s_v \rangle(k)$ between innovator degree and average size of vulnerable trees in both data and model with the same r values as in (b). Calculations were carried out on networks with size $N = 10^6$ and were averaged over 10^2 realisations. This figure was published in [199].

a large component, for larger r cascades cannot evolve since there are not enough nodes to fulfil the threshold condition of susceptible stable nodes, even if the underlying network is still connected. The transition point between these two phases of the adoption network is located at $r_c^{casc} \leq r_c^{net}$, limited from above by the critical point r_c^{net}. Finally, we observe from the empirical data and model results that the adoption network is held together by a large connected component of stable nodes. Consequently, for increasing r the stable adoption network goes through a percolation transition as well, with a critical point $r_c^{stab} \leq r_c^{casc} \leq r_c^{net}$.

To characterise these percolation phase transitions we compute the average size of the largest (LC) and second largest (LC^{2nd}) connected components (Fig. 4.9). We measure these quantities for the underlying network, and for the stable, vulnerable and global adoption networks, as a function of the fraction of immune nodes r. After $T = 89$ iterations (matching the length of the real observation period), we identify three regimes of the dynamics: if $0 < r < 0.6$ (dark-shaded area) the spreading process is very rapid and evolves as a global cascade, which reaches most of the nodes of the shrinking susceptible network in a few iteration steps. About 10% of adopters are connected in a percolating stable cluster, while vulnerable components remain very small in accordance with empirical observations. In the crossover regime $0.6 < r < 0.8$ (light-shaded area), the adoption process slows down considerably (Fig. 4.9, upper panel), as stable adoptions become less likely due to the quenching effect of immune nodes. The adoption process becomes the slowest at $r_c^{stab} = 0.8$ when the percolating stable cluster falls apart, as demonstrated by a peak in the corresponding LC^{2nd} curve in Fig. 4.9 (diamonds in lower panel). Finally, around $r_c^{casc} = 0.9$ the adoption network becomes fragmented and no global cascade takes place. Since the underlying network has a broad degree distribution, it is robust against random node removal processes [278]. That is why its critical percolation point r_c^{net} appears after 95% or more nodes are immune. Note that similar calculations for another service have been presented [199] with qualitatively the same results, but with the crossover regime shifted towards larger r due to different parameter values of the model process.

We can use these calculations to estimate the only unknown parameter, namely the fraction r of immune nodes in Skype, by matching the relative size of the largest component (LC_{Net})

between real and model adoption networks at time T. Empirically, this value is the relative size $s_a^{LC} \simeq 0.043$ (for more details see [199]). Matching this relative size with the simulation results (see the observation line in Fig. 4.9a upper panel), we find that it corresponds to $r^{emp} = 0.73$ (prediction line in Fig. 4.9a), suggesting that the real adoption process lies in the crossover regime. In other words, large adoption cascades could potentially evolve in Skype but with reduced speed, as 73% of users might not be interested in adopting a service within the network.

To test the validity of the predicted r^{emp} value we perform three different calculations. First we measure the maximum relative growth rate of cumulative adoptions and find a good match between model and data (see Skype s3 and Model Skype s3 in Fig. 4.2a). In other words, the model correctly estimates the speed of the adoption process. Second, we measure the distribution $P(d)$ of the depths of induced vulnerable trees (Fig. 4.9b). Vulnerable trees evolve with a shallow structure in the empirical and model processes. After measuring the distribution $P(d)$ for various r values below, above and at r^{emp}, we find that the distribution corresponding to the predicted r^{emp} value fits the best with the empirical data. Finally, in order to verify earlier theoretical suggestions [324], we look at the correlation $\langle s_v \rangle(k)$ between the degree of innovators and the average size of vulnerable trees induced by them (Fig. 4.9c). Similar to the distribution $P(d)$, we perform this measurement on the real data and in the model for $r = 0.6$ and 0.9, as well as for the predicted value $r_{emp} = 0.73$. We find a strong positive correlation in the data, explained partially by degree heterogeneities in the underlying social network, but surprisingly well emulated by the model as well. More importantly, although this quantity appears to scale with r, the estimated r value fits the empirical data remarkably well, thus validating our estimation method for r based on a matching of relative component sizes.

4.5.3 Threshold driven contagion on weighted networks

Weighted networks provide meaningful representations of the architecture of a large number of complex systems where weights capture the strength of interactions between connected entities. This way, weights can help to differentiate between links of varying importance, influence, and role, which may play crucial roles in dynamical processes evolving on networks. In threshold models on networks, links are usually considered unweighted, such that the stimuli or influence arriving from each neighbour contributes equally to reaching the behavioural threshold. Although this assumption simplifies their modelling, it does not lead to an accurate representation of real world dynamics. For example, in neural systems synaptic connections have weights that quantify the strength of incoming stimuli, and contribute unequally in bringing neurons to an excited state, as recognised recently in models of neural population dynamics [177]. In social systems link weights are associated with tie strengths that quantify the social influence that individuals have on their peers. Measurement of tie strength is a long standing challenge, but it is generally accepted that social ties are not equal, as some are more influential than others on one's decision making. Surprisingly, apart from some recent studies [210, 175, 94], weights have been commonly overlooked in models of threshold driven phenomena. In the next Section, we summarise our work on threshold driven contagion on weighted networks. This study is a natural extension of our previously discussed dynamical threshold model, but it sheds the light on more exotic, not yet observed dynamical behaviour. Results have been published originally in [354], while here we present the essence of our findings only.

To study threshold driven dynamical processes over weighted networks we build on the model defined in the previous Section and in [316, 199]. Following its standard formulation [368, 324, 316, 199], we define a monotone binary-state dynamics over a weighted, undirected network where edge weights $w > 0$ are continuous variables drawn from the distribution $P(w)$. The edge weight w_{ij} represents the capacity of connected nodes i and j to influence each other. Accordingly, the

node strength $q_k(i) = \sum_{j=1}^{k} w_{ij}$ is the total influence what node i receives from its k neighbours. However, influence may vary from neighbour to neighbour. We implement this idea by defining the partial strength $q_m(i) = \sum_{j=1}^{m} w_{ij}$ associated with the influence of the m infected neighbours on node i (where $0 \leq m \leq k$). If the condition $q_m \geq \varphi q_k$ is fulfilled, node i becomes infected and remains so indefinitely. For simplicity we assume that all nodes have the same threshold $\phi = \varphi$, just as in many other studies [368, 324].

Similar as earlier, we build our analytical description on Gleeson's approximate master equation formalism for stochastic binary-state dynamics [302, 140, 138, 139]. In order to extend this formalism to weighted networks, we discretise $P(w)$ and assume only n possible weight types w_j, such that all distinct weights in the network are contained in the weight vector $\mathbf{w} = (w_1, \ldots, w_n)^{\mathrm{T}}$. Then, a node in class (k,m) has k_j links with weight w_j and $m_j = 0, \ldots, k_j$ infected neighbours across these links, such that $k = \sum_{j=1}^{n} k_j$ and $m = \sum_{j=1}^{n} m_j$. Furthermore, we can define a degree vector $\mathbf{k} = (k_1, \ldots, k_n)^{\mathrm{T}}$ and a partial degree vector $\mathbf{m} = (m_1, \ldots, m_n)^{\mathrm{T}}$, generalising the strength and partial strength to $q_{\mathbf{k}} = \mathbf{k} \cdot \mathbf{w}$ and $q_{\mathbf{m}} = \mathbf{m} \cdot \mathbf{w}$, respectively. Nodes in class $(\mathbf{k}, \mathbf{m})$ have identical strengths and partial strengths, and follow the same pair of rate equations for the fraction $s_{\mathbf{k},\mathbf{m}}(t)$ (resp. $i_{\mathbf{k},\mathbf{m}}(t)$) of $\mathbf{k}$-nodes that are susceptible (resp. infected) at time t and have partial degree vector $\mathbf{m}$.

In our threshold driven model, a susceptible node can become infected in two ways, either spontaneously with rate p, or if its weighted threshold ϕ is reached. As such, the infection rate of susceptible nodes in class $(\mathbf{k}, \mathbf{m})$ is

$$F_{\mathbf{k},\mathbf{m}}\mathbf{k} = \begin{cases} p & q_{\mathbf{m}} < \phi q_{\mathbf{k}} \\ 1 & q_{\mathbf{m}} \geq \phi q_{\mathbf{k}} \end{cases}, \quad k > 0, \tag{4.23}$$

with $F_{\mathbf{0},\mathbf{0}} = p$. The stepwise nature of $F_{\mathbf{k},\mathbf{m}}\mathbf{k}$ allows us to map the rate equations for $s_{\mathbf{k},\mathbf{m}}$ and $i_{\mathbf{k},\mathbf{m}}$ to a reduced-dimension system, as has been done earlier in Eq.4.24 for unweighted complex contagion [316, 199]. Namely, if we consider as aggregated variables the density $\rho(t)$ of infected nodes and the probability $\nu_j(t)$ that a randomly chosen neighbour (across a j-type edge) of a susceptible node is infected, then the description of the dynamics can be reduced to the system of $n+1$ equations

$$\dot{\nu}_j = g_j(\nu, t) - \nu_j, \tag{4.24a}$$

$$\dot{\rho} = h(\nu, t) - \rho, \tag{4.24b}$$

where $\nu = (\nu_1, \ldots, \nu_n)^{\mathrm{T}}$ is the vector of probabilities ν_j for all weight types, and $g_j(\nu, t)$ and $h(\nu, t)$ are functions of binomial terms.

Regular networks with bimodal weights

To study the dynamics of our model we first consider a simple structure, the configuration-model k-regular network, with $k = 7$. Edge weights are sampled from a bimodal distribution with $n = 2$ values, denoted strong (w_1) and weak (w_2). The weight distribution is characterised by its average μ, standard deviation $\sigma \geq 0$, and the fraction δ of links that are strong. Thus, weights take the values $w_1 = \mu + \sigma\sqrt{(1-\delta)/\delta}$ and $w_2 = \mu - \sigma\sqrt{\delta/(1-\delta)}$. The parameter δ contributes to the skewness of $P(w)$, initially fixed to the symmetric case $\delta = 0.5$. The parameter σ interpolates weight heterogeneity between the homogeneous case of an unweighted network ($\sigma = 0$), and the most heterogeneous case of a diluted network ($\sigma = \mu\sqrt{(1-\delta)/\delta}$), where only strong links have influence and the weak are functionally absent. After fixing the spontaneous infection rate p and skewness δ, our model has only two parameters, σ and ϕ. To characterise the speed of dynamics we introduce the quantity t_a, the time when infection density reaches a set value ($\rho = 0.75$), called the absolute time of cascade emergence. We measure t_a via numerical simulations of the (σ, ϕ)-parameter space, which shows unexpected dependencies on both parameters. On one hand, for

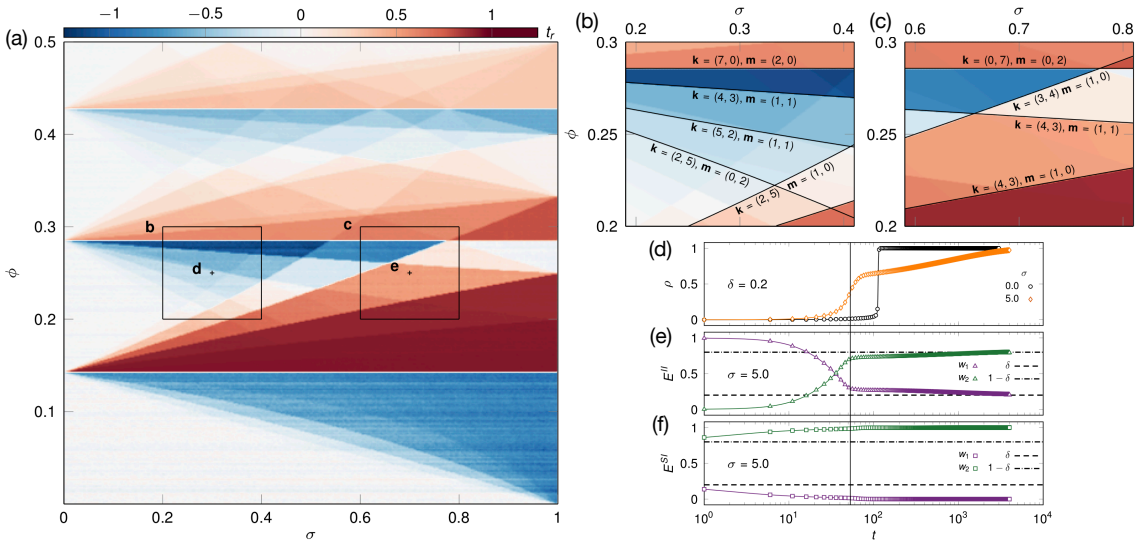


Figure 4.10: Relative time of threshold driven cascades on weighted networks. (a) Relative time t_r of cascade emergence on (σ,ϕ)-parameter space, simulated over k-regular regular networks ($k=7$) with $\mu=1$, $\delta=0.5$, $p=2\times10^{-4}$, $N=10^4$ and averaged over 25 realisations. (b-c) Selected regions of parameter space in (a), where t_r is instead calculated from the numerical solution of the AME systems in Eq. 4.24. Boundaries are obtained from a combinatorial argument [354] for various $(\mathbf{k},\mathbf{m})$ classes. (d-f) Effect of skewed weight distributions on cascade evolution. (d) Infection density $\rho(t)$ on k-regular networks ($k=7$) and a bimodal weight distribution with $\mu=3$ and $\delta=0.2$, both for unweighted ($\sigma=0$) and heterogeneous ($\sigma>0$) cases. (d-e) Fractions of strong (w_1) and weak (w_2) links connecting two infected nodes in the bulk of the infected component [$E^{II}(t)$, b] and susceptible and infected nodes on its surface [$E^{SI}(t)$, c] in the heterogeneous spreading scenario of (a). Simulations (symbols) are averaged over 25 realisations with $p=2\times10^{-4}$ and $N=10^4$, and compared with the corresponding AME solutions (lines). Dashed lines are the expected fractions of weak and strong links as determined by δ, and the vertical line shows the inflection point of ρ in the heterogeneous case of (a), which coincides with a turning point of E^{II} in (b). This figure was prepared by S. Unicomb and was published in [354].

fixed σ and increasing ϕ the dynamics slows down, since nodes with higher thresholds require more infected neighbours to become infected. On the other, for fixed ϕ the dynamics depends *non-monotonously* on σ, where cascades may evolve either faster or slower as we increase weight heterogeneity, relative to the unweighted case ($\sigma=0$).

To better characterize the dynamics, first we concentrate on the σ dependency by calculating $t_r=[t_a(0,\phi)-t_a(\sigma,\phi)]/t_a(0,\phi)$, the time of cascade emergence relative to the unweighted case with the same ϕ value. (Fig. 4.10a). The relative time t_r will be positive if the weighted process evolves faster than the unweighted case, zero if they evolve at the same speed, and negative if slower than the unweighted case. The (σ,ϕ)-parameter space for t_r is highly structured and driven by competing effects of key $(\mathbf{k},\mathbf{m})$ classes, which either reduce or enhance the speed of the spreading process as compared to the unweighted case. We also explore the corresponding numerical solution of the AME systems in Eq. 4.24, as well as an independent combinatorial solution [354] for the boundaries between regions of low and high cascade speed (Fig. 4.10b-c). Both the AME and combinatorial solutions perfectly recover the parameter space obtained by simulations. To further explore how weight heterogeneities produce slow or fast cascades, we partition the system according to the number m of infected neighbours required for infection, and measure the aggregated infection rate $F_{k,m}(t)=\sum_{\mathbf{k},\mathbf{m}}P(\mathbf{k})F_{\mathbf{k},\mathbf{m}}s_{\mathbf{k},\mathbf{m}}(t)/\sum_{\mathbf{k},\mathbf{m}}P(\mathbf{k})s_{\mathbf{k},\mathbf{m}}(t)$ and other determinant quantities in several spreading scenarios (Fig. 4.10d-e).

In the neutral scenario, all $(\mathbf{k},\mathbf{m})$ classes of the weighted network share the same dynamics as the corresponding (k,m) class in an unweighted network, so $F_{k,m}=p$ or 1 and weights have

no impact on contagion, meaning $t_r = 0$. In a decelerative scenario like $\phi = 0.25$ and $\sigma = 0.3$ (Fig. 4.10d), $F_{k,m}$ for any m is equal to its unweighted counterpart, except for the $m = 2$ class. Here, the adoption rate is 1 in the unweighted case but strongly suppressed in the weighted case, thus decreasing the overall spreading speed. For an accelerative scenario, like $\phi = 0.25$ and $\sigma = 0.7$, competing effects from several $(\mathbf{k}, \mathbf{m})$ classes combine to determine the overall dynamics (Fig. 4.10e). The rate $F_{k,m}$ for $m = 2, \ldots, 4$ is lower than 1 which is a decelerative effect (as in the previous case), but the rate $F_{k,1}$, which is equal to p in the unweighted case, is significantly larger than p here. Since at the early stages of contagion the number of nodes in class $m = 1$ is larger than in any other class with $m > 1$, spreading evolves rapidly to an early cascade. It should be noted that competition between the accelerative and decelerative effects of the weight distribution is one of the defining characteristics of threshold driven contagion on weighted networks. It is this competition that leads to the interference patterns evident in Fig. 4.10a.

Up until now we have considered the symmetric case $\delta = 0.5$ with equal numbers of strong and weak links. However, by skewing the weight distribution we observe an additional effect of weight heterogeneities on the spreading behaviour. When $\delta = 0.2$ the extent of the cascade decreases for large σ with respect to the unweighted case (Fig. 4.10d). In this case, despite their sparsity, strong links again drive the contagion, but are soon exhausted causing spreading to slow down and continue via spontaneous or infrequent threshold driven infections over weak ties (Fig. 4.10e). Indeed, strong links dominate the bulk of the infected component, but disappear quickly from its surface (Fig. 4.10f). These so-called *partial cascades*, which do not infect the whole system through the cascade, are associated with skewness and a sufficiently large standard deviation in the weight distribution and are reminiscent of the slow spreading caused by immune nodes, as well as low connectivity networks in unweighted complex contagion [316, 199, 368]. Overall, we identify non-monotonous spreading behaviour and partial cascades as the main consequences of weight heterogeneities in threshold driven contagion.

Although regular networks and bimodal weights are useful in characterising the qualitative impact of weights on contagion, they are rather unrealistic since real complex networks commonly appear with broad degree and weight distributions [42]. For this reason we further explored (not shown here [354]) the effects of heterogeneous degree and weight distributions on the threshold driven contagion. Interestingly, it turned out that the observed non-monotonous dependency of cascade time is robust against such heterogeneities, and were found even in case of data-driven simulations, where the contagion was iterated on real weighted networks.

To summarise, in this study we explored weighted networks with increasing complexity, from configuration-model structures with bimodal or lognormal weights, to real world networks with broad degree and weight distributions. We showed that threshold driven contagion depends non-monotonously on weight heterogeneity, creating slow or fast cascades relative to the equivalent unweighted spreading process. Via numerical simulations, master equations and combinatorial arguments, we found that this effect is the result of competing configurations of degree, weight, and infected neighbours that slow down or speed up contagion. We also observed that an imbalance in the amount of large and small weights leads to partial cascades, and smoother temporal patterns of spreading than those in unweighted networks. By analysing a range of degree and weight configurations, we show that these features are systemic and thus may drive a variety of real world contagion phenomena.

4.5.4 Threshold driven contagion on multiplex networks

In earlier studies of complex contagion [152, 368, 316, 189, 316], cascades were predicted in single layer sparse networks. In these structures nodes typically have smaller degrees making them easier to reach the necessary threshold of relative neighbor influence to join a spreading cascade.

In empirical social networks, however, individuals can maintain hundreds of ties [147, 112], with interaction strength varying across social contexts [183, 354, 70], yet still exhibit frequent system-wide cascades of social contagion [199, 109, 142]. To address this puzzle we incorporated additional relevant features motivated from empirical social networks into our earlier defined threshold model [354]. We consider that edges in a social network to vary in "quality" [155, 208], usually associated with the intimacy or perceived importance of a relationship between individuals [391], and scale with the strength of interpersonal influence [81, 350]. Heterogeneity in tie quality is well modeled by multiplex structures [216, 55], particularly in case of social contagion [383, 68, 235, 392]. In multiplex models of social networks, individual layers represent the social context of a relationship (e.g. kinship, acquaintance), allowing us to classify ties by social closeness, as recognised by Dunbar's intimacy circle theory [391] (for more details see 1.2.2). According to this theory ego networks comprise a small but *high-intimacy* circle of close relationships, like family and long term friends, followed by large but *low-intimacy* circles of distant friends and acquaintances.

Motivated by these observations, we defined in [356] a contagion as a binary-state dynamics over a weighted, undirected multiplex network of N nodes connected throughout M layers (see for schematic demonstration 4.11a-d). A node represents an individual u, and layer i the social context in which individuals interact, $1 \leq i \leq M$. The degree of u in each layer i takes discrete values $k_i = 0,\ldots,N-1$ according to the degree distribution $P_i(k)$. Edge weights $w_i(u,v)$ follow the continuous distribution $P_i(w)$ and capture the total capacity of nodes u and v to influence each other via layer i. The network allows for layer overlap [74] as nodes may be connected in multiple layers, modelling individuals who share several social contexts (see 4.11c). For simplicity, we assume that node degree is independent across layers, and that degree and weight distributions $P_i(k)$ and $P_i(w)$ differ by layer explicitly in their means $z_i = \sum_k kP_i(k)$ and $w_i = \int wP_i(w)dw$, but otherwise retaining their functional form. In order to reproduce the hierarchical organisation of edges suggested by intimacy circle theory [391], we assume that the mean degree z_i and weight w_i scale with layer index i as

$$z_{i+1} = \delta_z z_i \quad \text{and} \quad w_{i+1} = \delta_w w_i, \tag{4.25}$$

with $\delta_z \geq 1$ and $\delta_w \leq 1$. In other words, ego networks comprise a small number of high-intimacy neighbours [4.11(a)] and a larger number of low-intimacy neighbours (see 4.11(b)). For more details on the model setting see [356].

Since nodes must be either connected or disconnected via each of the M network layers, their interaction is characterised by one of $2^M - 1$ resultant edge types (see 4.11(d)), disregarding nodes disconnected in all layers, and indexing by j such that $1 \leq j \leq 2^M - 1$. Node configuration is thus described by the number of neighbours k_j and infected neighbours m_j across edges of type j, with $0 \leq m_j \leq k_j$. We store k_j and m_j in the degree vector $\mathbf{k}$ and partial degree vector $\mathbf{m}$, respectively (of dimension $2^M - 1$). Note that we consistently index layer by i and resultant edge type by j.

The threshold rule proposed by Watts [368] can be extended to multiplex networks in several ways. Denoting the set of neighbours of node u in layer i by $\mathcal{N}_i(u)$, the total influence upon u in layer i is $q_{k_i} = \sum_{v\in\mathcal{N}_i(u)} w_i(u,v)$. Restricted to infected neighbours, $\mathcal{N}_i(u)|_I$, this gives $q_{m_i} = \sum_{v\in\mathcal{N}_i(u)|_I} w_i(u,v)$. In one variant of the threshold rule, that we call *weighted sum*, rule nodes perceive influence in aggregate, summed over layers and adopt with respect to a single threshold if $q_m \geq \phi q_k$, where $q_k = \sum_i q_{k_i}$ and $q_m = \sum_i q_{m_i}$. Note that we studied two other threshold rules, called *multiplex AND* and *multiplex OR*, with results reported in [356].

Similar to the methodology discussed in Section 4.5.3, we solve our model using the approximate master equation formalism [139, 140, 316, 199, 354]. Just as in case of weighted networks discussed earlier, we derive the density of infected nodes ρ and the average probability ν_j that a j-type neighbor of a susceptible node is infected at time t in a similar reduced form as shown in Eq. 4.24a and b (resp.).

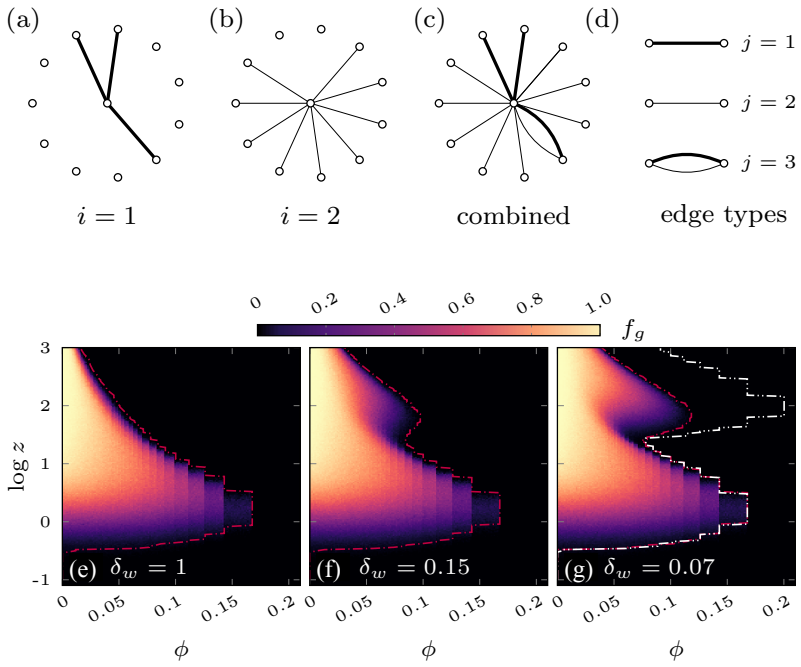


Figure 4.11: (a-b) Egocentric view of multiplex structure with $M = 2$ layers, where edge density increases ($\delta_z > 1$) and edge weight decreases ($\delta_w < 1$) in each layer i. (c) Egocentric network overlap between layers. (d) Emergent edge types in the overlapping network. In the multiplex, the central node has degree vector $\mathbf{k} = (2,8,1)^T$, encoding layer overlap. (e-g) Emergence of a high-z cascading phase in (ϕ, z)-space for the weighted sum rule, for log-normal degree distribution with fixed $\delta_z = 50$, $\gamma = 0.5$ and decreasing δ_w. The standard deviation of the LN distribution is $\sigma_{k_i} = 2z_i$. MC simulations provide the relative frequency f_g of global cascades, after 10^3 instances of single node perturbation, in a configuration-model multiplex with $N = 10^6$. In (e) we recover the classic Watts phase diagram ($\delta_w = 1$). The constraint $\langle w \rangle = 1$ means $\mathbf{w} = (1,1)^T$, $(6,0.9)^T$ and $(11,0.8)^T$, from (e) to (g). The outer contour (dash-double dotted white line) in (g) shows the case $\delta_w \to 0$ [$\delta_w = 10^{-3}$]. Dash-dotted red lines show agreement with linear stability analysis prediction.

A numerical solution of these equations (in [354]) provides the dynamical evolution of the system, while a linear stability analysis identifies the region in the (ϕ, z)-space, which allow global cascades (indicated as dash-dotted lines in Figs. 4.11e. We derive a global cascade condition via the Jacobian matrix $\mathbf{J}$ corresponding to 4.24, evaluated at the fixed point $\nu^* = 0$ to study the response of the network to an infinitesimal perturbation, or single infected seed. Using this method we compute the relative frequency f_g of global cascades from Monte Carlo simulations, as shown in Fig. 4.11e-g.

Surprisingly, the weighted sum rule leads to a high-z cascading phase, and thus to a re-entrant phase transitions for constant ϕ, in an $M = 2$ layer multiplex with a log-normal degree distribution in each layer as shown in 4.11g (for details see [356]). In two layers, we define layer overlap as $\gamma = |E_1 \cap E_2|/|E_1|$, where E_i is the edge set in layer $i = 1,2$ ($|E_1| < |E_2|$). We can increase weight heterogeneity by decreasing the weight scaling factor δ_w, resulting in a second cascading regime.

According to the Watts cascade model, in a single layer network cascading phase appears in (ϕ, z)-space when vulnerable nodes form a percolating cluster and the network is sparse. On the other hand, for large z results in most nodes being stable against neighbor infection, and cascades becoming exponentially rare. However, under the weighted sum rule, weight heterogeneity allows one high-influence infected neighbor to dominate a node's total received influence if remaining neighbours have low influence. Crucially, such configurations are abundant when the conditions

$\delta_z > 1$ and $\delta_w < 1$ are satisfied simultaneously, resulting in a percolating vulnerable cluster at high z. In other words, in the low-z phase, cascades are mediated by the connectivity of the weak layer, since the strong layer is too sparse to percolate. In the high-z phase, strong edges percolate and determine the stability of adjacent nodes that are otherwise stable to the dense weak layer.

As reported in [356], similar phenomena appears for the other thresholds too, even in empirical multilayer structures exported from Twitter. Moreover the number of re-entrant phase transitions are determined by the number of layers in the network.

4.5.5 Threshold driven contagion on temporal networks

In a recent study [353] we extended further our understanding of complex contagion processes on temporal networks and the effect of link burstiness on the final spreading outcome. Threshold driven dynamics over static networks have been extensively studied both empirically [199] and theoretically [368, 138, 199, 354, 356], but analysis of their behaviour on temporal networks is so far limited to a small number of empirical studies [192, 190, 337, 29]. Using random reference models of temporal networks, it has been shown that when infection is driven by the fraction of infected neighbours, rather than their absolute number, bursty interactions may lead to deceleration [192, 190, 29]. In contrast, if the threshold measure of influence is absolute, burstiness may have an accelerative effect [192]. Acceleration has been also observed in case of history-dependent contagion [337].

Another aspects of information diffusion in social and economic settings are the limits inherent to our social and cognitive capacities, namely that we have finite attention. A mechanism that has emerged to deal with these limitations is ephemeral content [51, 321] that disappears after a specified amount of time, in principle concentrating the attention of viewers. In contrast, persistent content is not explicitly erased, but due to cognitive limits and competing content, gradually decreases in visibility.

In [353], we propose an analytical framework to systematically describe the relationship between the diffusion of information, bursty temporal interactions, and inherent limits to our attention and memory, thus providing the theoretical foundation necessary to shed light on the role of burstiness in generic diffusion processes, including simple and complex contagion models of physical, biological and social phenomena.

More precisely, we incorporate the most widely documented features of temporal interactions into a framework of binary state dynamics and benchmark its behaviour with standard models of threshold driven and epidemic spreading. Although stochastic bursty interactions are likely emergent phenomena [39, 361], their dynamics are well approximated by renewal processes [377]. Temporal heterogeneity in network interactions can then be characterised by the variability in inter-event times τ, the time between consecutive events on a given edge, parameterised by the inter-event time distribution $P(\tau)$, while other features of the temporal network are considered maximally random. Renewal processes represent the simplest model of bursty, non-Markovian dynamics, and a departure from the memoryless assumption implicit in Poisson processes. Nevertheless, we are able to show that such a system can be accurately captured by a modified version of the approximate master equation (AME) formalism, which is essentially memoryless, implying the existence of a purely Markovian system with almost identical behaviour. We show both analytically and numerically that bursty temporal interactions give rise to a percolation transition in the connectivity of the temporal network, separating phases of slow and rapid dynamics for both epidemic and threshold models of information diffusion. We find that diffusion dynamics are sensitive to the choice of inter-event time distribution, particularly in regard to its skewness, and we demonstrate a data collapse across distributions when controlling for this effect.

* * *

Finally, in a separate work [199], we proposed another model of complex contagion phenomena, which build on an SIR model scheme, but relying on the observations reported in Section 4.3.1. It extends the model with a linearly increasing influence of adoption with the fraction of adopting neighbours.

4.6 Conclusion

This Chapter focused on my and my co-authors contributions on the observation and modelling of collective social phenomena. After a brief introduction I summarised selected studies on static and dynamic observations of collective behaviour, together with statistical and mechanistic models built on the obtained observations. I payed special focus on spreading processes, by walking through studies considering simple and complex contagion phenomena. My aim in this Chapter was to give an overview about my motivations and main contributions to the understanding of collective social behaviour. This way I decided to exclude several papers from the detailed discussion, but mentioning them in short summaries in the end of each Section.

5. Discussion

5.1 Perspectives

The potential future of my research and my field can be discussed at different time-scales. Here I am going to describe some immediate and midterm directions related to my own research, while I will also synthesise my view on the long-term future potentials of my associated fields.

5.1.1 Future directions of my research

With my collaborators we have continued several research lines summarised in this Thesis, while we also initiated new spin-off directions based on recent ideas and developments. Next I shortly describe the most promising angles.

Operative and fundamental COVID-19 research

Since March 2020, from the very beginning of the COVID-19 pandemic in Hungary I got involved in the mathematical modelling force-team developing predictive epidemic models for decision makers. Within this larger operations I have been the responsible for a longitudinal data collection effort aiming to record age stratified contact matrices dynamically using online and representative phone survey methods. While the phone survey data collection records the answers of $1000-1500$ respondents, the online questionnaire reached over $230,000$ people responding more than $480,000$ times [204]. We exploited this data collection infrastructure to collect dynamical data on behavioural changes related to the actual pandemic regarding vaccination and self-protection attitudes, danger awareness and perception, socioeconomic status, work conditions, mobility etc. Using this data I started several new research directions with colleagues from various fields. Example are the development of age stratified contact matrices reconstruction methods [221], the study of switchover phenomena on metapopulation networks [285], or a study disclosing vaccination preferences in Hungary [229]. On the longer term I plan to study distance dependent awareness modelling and its effect on epidemic processes and also metapopulation models of epidemic incorporating not only age but sociodemographically and economically stratified contact matrices.

Social experiments

Together with some socio-linguist colleagues we recently completed a large-scale social experiment [97], where we collect social and verbal interactions of children between age in 3 and 6. Our experiment employs RFID sensors, which are capable of recording face-to-face interactions and speech at the same time, and is accompanied with frequent questionnaires about the socioeconomic background and linguistic development of the participants. This project was running over three years by recording one week of data every months with the involvement of 110 children. Our scientific goal is to understand the emergent consensus on the usage of standard patterns of French language as the function of the social network, time, linguistic and socioeconomic background.

Data for Good

In another line of research we continue to analyse large-scale datasets recording the social networks and sociodemographic, location, and linguistic characters of individuals using data from online social systems, mobile communication, and satellites, and census. Our overall goal is to better understand the emergence of socioeconomic inequalities, limited social mobility, linguistic variance, and to infer average socioeconomic status at any level of granularity. I have several results in this direction, which were only mentioned in this thesis in Section 4.2, yet this is one of the main directions I would like to explore in the future.

Higher order correlations in networks

The study on weighted event graph representation of temporal networks we reported in Section 3.4.2 is only a first step in this direction. As a next step we are working on its extension to explore higher-order temporal-structural patterns, to more precisely model spreading processes on temporal networks, to identify large-scale temporal motifs, to define computationally cheap methods to calculate temporal network centralities, and to use this way of representation in systems where temporal interactions are important like in neural networks, transportation, or communication networks.

Rich datasets describing the actions and interactions of individuals provide outstanding sources of information and knowledge and fuel a wide spectrum of data-driven numerical simulations of dynamical processes. Data alone, however, even in huge amounts, do not easily transform into knowledge or predictive models. The rich and diverse information they contain raises crucial challenges concerning their analysis, representation and interpretation, the extraction of meaningful structures, and their integration into data-driven models. In a recent project we aim to build a methodological framework to reduce networked data complexity, while preserving its richness, by working at intermediate scales ("mesoscales"). Our objective is to reach a theoretical understanding and representation of rich and complex networked datasets for their use in predictive data-driven models.

In other exploratory projects we develop unsupervised learning approaches to build network embeddings (a) for static networks to infer correlations/patterns between node features and the mesoscopic structure of networks [241], and (b) for temporal networks to find correlations between events and infected populations induced by a spreading initiated from the actual event.

Modelling collective phenomena

In terms of social contagion phenomena one of my oldest personal puzzle is related to the phenomenological differences between simple and complex contagion processes. At the observation level these two processes appear very similar in real settings thus it is difficult to identify which of them is driving the actual spreading process. As both simple and complex mechanisms are arguably present simultaneously for a single social contagion process, beyond the identification of one or the other, it is important to decide which one plays a dominant role during the adoption of a single individual. Solving this problem would (a) close a long lasting debate about the dominant

mechanisms driving social contagion and (b) would bring us closer to understand the role of social influence during decision making, which would contribute to another historic open question about the role of influence vs. homophily in network formation.

5.1.2 Perspectives of my associated fields

Social physics and computational social science builds on decades of knowledge accumulated in conventional social sciences, but as a separate field it is relatively young [234]. First of all, its innovation is methodological as it relies on large and automatically collected behavioural data and novel experimental opportunities provided by the digital age. This new era just started and delivers new datasets, collection methods and challenges in a rapid pace. These developments fuel quantitative studies of social systems to identify mechanisms and emergent patterns in individual and social behaviour, with the promise of better understanding and predictability of human behaviour.

All three research domains I discussed in this Thesis have some relevance in these challenges, however, they are at different phases of their developments. The observation of bursty human behaviour is relatively recent [39], but was followed by a very intense period of investigations [200], which brought us to the point that its first-order characterisation is well developed. New questions in this direction are related to higher-order temporal and structural correlations [320, 230, 217, 49], which would bring us closer to understand the possible explanations and various consequences of such behavioural patterns. Temporal networks is a field which was also fuelled by the emergence of recently collected datasets recording time-varying interactions of entities on a high temporal resolution. This field has reached its first milestones [173, 170] but still offers several open theoretical and methodological challenges. Most interesting questions to answer are related to a more formal foundation of the representation, analytical tools and computability of temporal networks, in addition to the development of mathematically treatable network models. In addition, coupled dynamics of networks and processes are also very much un-understood. Due to the broad relevance of temporal networks in several fields, in my opinion, this field will remain in the centre of network science for a long period. Finally, the observation and modelling of the emergence of collective phenomena is a historic challenge, with some recent developments, which introduce even more realistic data-driven prediction of collective processes. The observation of global information diffusion and formation of collective opinion, the precise data-driven simulations of epidemic processes, or the observation and prediction of the mobility of millions are all promising advances which will carry this field on to explore and explain more complex phenomena.

5.2 Conclusions

While writing this Thesis I had two main goals in mind. On one hand, I wanted to write a concise summary of my most interesting scientific contributions, and on the other hand, I aimed to provide and up-to-date view and perspectives about my field. Due to my training and scientific interests, this Thesis emerged as an interdisciplinary work discussing topics and methods associated to physics, computer science, statistics, applied mathematics and computational social science. As a consequence, it is not easy to read with an approach from a single discipline. To help the reader on this end, beyond the technical in-depth discussion of the actual methodologies and findings, I was trying to design a line of description, which is accessible for a broader audience. Despite the diverse subjects of my works, I needed to identify a common ground bringing them all together under the hood of a single thesis. As an outcome, I entitled my thesis as *computational human dynamics*, which precisely captured the common denominator of all my contributions. In accordance, I consciously built all my reasoning and motivations on examples borrowed from social phenomena. On the other, some of my methodological work may be relevant in other fields as neural science or system biology, what I emphasised on the run.

I started my dissertation with an introduction to position the reader on the landscape of my field, open up actual challenges, and to put in perspective my contributions. To purposefully ground the terminology used throughout the whole dissertation, I briefly introduced the reader to the basic concepts of complex networks and to the characteristics and theories of social networks. To avoid redundant descriptions and confusing cross-referencing, I described various datasets, which had been used on multiple occasions in the reported studies. Subsequently, I summarised my contributions in three scientific domains. First, I concentrated on the bursty, heterogeneous temporal characters of individual human dynamics. Recently written a book on this topic, I had the advantage to easily provide an overall view on this field, and to introduce my contributions about the observations, characterisation and modelling of *bursty phenomena*. Next, I concentrated on my works related to *temporal networks*. This Chapter synthesises my publications I co-authored over the last years on various novel methods for the representation, characterisation, and modelling of time-varying structures. Our contributions range from system level characterisation and random reference models to the detection of higher-order correlations and the development of activity-driven generative network models of temporal networks. Finally, as a third area of my contributions, I discussed my work on the data-driven observations and modelling of *collective social phenomena*. First I summarised studies on the static and dynamical observations of emergent patterns of collective processes. Partially building on these observations, I subsequently discussed the modelling of simple and contagion phenomena, to understand how to control them on temporal structure, how to predict their dynamical emergence on society-large networks with various heterogeneous characters.

This Thesis was written for the purpose of obtaining the degree of Doctor of Science of the Hungarian Academy of Sciences. However, it was meant for something more than being simply a summary of my contributions. I wrote it as a research statement and a milestone (for myself) to summarise my past achievements, to clarify my actual position, and to help building plans for my future research. I hope that some others will also find this synthesis useful and motivating to find new ideas to solve and to explore paths towards a more comprehensive understanding the world around.

Bibliography

[1] Amazon Mechanical Turk. https://www.mturk.com. Accessed: 2018-01-17.

[2] arXiv.org. https://arxiv.org. Accessed: 2018-01-17.

[3] Dylnet. https://dylnet.univ-grenoble-alpes.fr/?language=en. Accessed: 2022-01-26.

[4] General data protection regulation, regulation 2016/679 by the european comission.

[5] Hungarian data supply questionnaire (maszk) (date of access 2020.09.28).

[6] Hungarian data supply questionnaire (maszk) team, https://covid.sed.hu/tabs/staff, (date of access 2020.09.28).

[7] Nemzeti adatvédelmi és információszabadság hatóság, https://www.naih.hu (date of access 2020.12.).

[8] SigKDD. http://www.kdd.org. Accessed: 2018-01-17.

[9] SurveyMonkey. https://www.surveymonkey.com. Accessed: 2018-01-17.

[10] Internet world stats, 2011. Accessed: 2013.07.15.

[11] Hungarian microcensus. https://www.ksh.hu/mikrocenzus2016/?lang=en, 2016.

[12] Bureau of transportation statistics, 2017.

[13] Central hungary, wikipedia, 2021. https://en.wikipedia.org/wiki/Central_Hungary [date of access: 17/05/2021].

[14] Jacob Levy Abitbol and Marton Karsai. Interpretable socioeconomic status inference from aerial imagery through urban patterns. *Nature Machine Intelligence*, 2(11):684–692, 2020.

[15] Jacob Levy Abitbol, Márton Karsai, Jean-Philippe Magué, Jean-Pierre Chevrot, and Eric Fleury. Socioeconomic dependencies of linguistic patterns in twitter: A multivariate analysis. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web*, pages 1125–1134. International World Wide Web Conferences Steering Committee, 2018.

[16] Amnon Aharony and Dietrich Stauffer. *Introduction to percolation theory*. Taylor & Francis, 2003.

[17] Réka Albert and Albert-László Barabási. Topology of evolving networks: local events and universality. *Physical review letters*, 85(24):5234, 2000.

[18] Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Reviews of modern physics*, 74(1):47, 2002.

[19] Talayeh Aledavood, Eduardo López, Sam GB Roberts, Felix Reed-Tsochas, Esteban Moro, Robin IM Dunbar, and Jari Saramäki. Daily rhythms in mobile telephone communication. *PloS one*, 10(9):e0138098, 2015.

[20] Laura Alessandretti, Márton Karsai, and Laetitia Gauvin. User-based representation of time-resolved multimodal public transportation networks. *Royal Society open science*, 3(7):160156, 2016.

[21] Paolo Allegrini, Danilo Menicucci, Remo Bedini, Leone Fronzoni, Angelo Gemignani, Paolo Grigolini, Bruce J. West, and Paolo Paradisi. Spontaneous brain activity as a source of ideal $1/f$ noise. *Phys. Rev. E*, 80:061914+, December 2009.

[22] Eduardo G. Altmann, Janet B. Pierrehumbert, and Adilson E. Motter. Beyond word frequency: Bursts, lulls, and scaling in the temporal distributions of words. *PLoS ONE*, 4(11):e7678+, November 2009.

[23] Kartik Anand and Ginestra Bianconi. Entropy measures for networks: Toward an information theory of complex topologies. *Physical Review E*, 80(4):045102, 2009.

[24] C Anteneodo, R Dean Malmgren, and DR Chialvo. Poissonian bursts in e-mail correspondence. *The European Physical Journal B*, 75(3):389–394, 2010.

[25] Andrea Apolloni, Chiara Poletto, José J Ramasco, Pablo Jensen, and Vittoria Colizza. Metapopulation epidemic models with heterogeneous mixing and travel behaviour. *Theoretical Biology and Medical Modelling*, 11(1):1–26, 2014.

[26] Sinan Aral, Lev Muchnik, and Arun Sundararajan. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proceedings of the National Academy of Sciences*, 106(51):21544–21549, 2009.

[27] Robert Axelrod. The dissemination of culture: A model with local convergence and global polarization. *Journal of conflict resolution*, 41(2):203–226, 1997.

[28] D. Ayala. *Temporal percolation in the Erdös-Rényi model and the effect of burstiness*. PhD thesis, University of Oxford, 2015.

[29] Ville-Pekka Backlund, Jari Saramäki, and Raj Kumar Pan. Effects of temporal correlations on cascades: Threshold models on temporal networks. *Physical Review E*, 89(6):062815, 2014.

[30] Lars Backstrom, Dan Huttenlocher, Jon Kleinberg, and Xiangyang Lan. Group formation in large social networks: membership, growth, and evolution. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 44–54. ACM, 2006.

[31] Arash Badie-Modiri, Márton Karsai, and Mikko Kivelä. Efficient limited-time reachability estimation in temporal networks. *Physical Review E*, 101(5):052303, 2020.

[32] Arash Badie-Modiri, Abbas K Rizi, Márton Karsai, and Mikko Kivelä. Directed percolation in random temporal network models with heterogeneities. *Physical Review E*, 105(5):054313, 2022.

[33] Arash Badie-Modiri, Abbas K Rizi, Márton Karsai, and Mikko Kivelä. Directed percolation in temporal networks. *Physical Review Research*, 4(2):L022047, 2022.

[34] Paolo Bajardi, Chiara Poletto, Jose J Ramasco, Michele Tizzoni, Vittoria Colizza, and Alessandro Vespignani. Human mobility networks, travel restrictions, and the global spread of 2009 h1n1 pandemic. *PloS one*, 6(1):e16591, 2011.

[35] Per Bak. *How nature works : the science of self-organized criticality*. Copernicus, 1996.

[36] Per Bak, Kim Christensen, Leon Danon, and Tim Scanlon. Unified Scaling Law for Earthquakes. *Phys. Rev. Lett.*, 88:178501, 2002.

[37] Eytan Bakshy, Itamar Rosenn, Cameron Marlow, and Lada Adamic. The role of social networks in information diffusion. In *Proceedings of the 21st international conference on World Wide Web*, pages 519–528. ACM, 2012.

[38] Raquel A Baños, Javier Borge-Holthoefer, and Yamir Moreno. The role of hidden influentials in the diffusion of online information cascades. *EPJ Data Science*, 2(1):6, 2013.

[39] Albert-László Barabási. The origin of bursts and heavy tails in human dynamics. *Nature*, 435(7039):207–211, 2005.

[40] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.

[41] Vladimir Barash. *The Dynamics of Social Contagion*. PhD thesis, Cornell University, Ithaca, NY, USA, 2011. AAI3485091.

[42] Alain Barrat, Marc Barthelemy, Romualdo Pastor-Satorras, and Alessandro Vespignani. The architecture of complex weighted networks. *Proceedings of the national academy of sciences*, 101(11):3747–3752, 2004.

[43] Alain Barrat, Marc Barthelemy, and Alessandro Vespignani. *Dynamical processes on complex networks*. Cambridge university press, 2008.

[44] Alain Barrat, Ciro Cattuto, Vittoria Colizza, Jean-François Pinton, Wouter Van den Broeck, and Alessandro Vespignani. High resolution dynamical mapping of social interactions with active rfid. *arXiv preprint arXiv:0811.4170*, 2008.

[45] Louise Barrett, Robin Dunbar, and John Lycett. *Human evolutionary psychology*. Princeton University Press, 2002.

[46] Marc Barthélemy. Spatial networks. *Physics Reports*, 499(1):1–101, 2011.

[47] Marc Barthélemy, Claude Godreche, and Jean-Marc Luck. Fluctuation effects in metapopulation models: percolation and pandemic threshold. *Journal of theoretical biology*, 267(4):554–564, 2010.

[48] Frank M Bass. A new product growth for model consumer durables. *Management science*, 15(5):215–227, 1969.

[49] Federico Battiston, Giulia Cencetti, Iacopo Iacopini, Vito Latora, Maxime Lucas, Alice Patania, Jean-Gabriel Young, and Giovanni Petri. Networks beyond pairwise interactions: structure and dynamics. *Physics Reports*, 874:1–92, 2020.

[50] Alex Bavelas. Communication patterns in task-oriented groups. *The Journal of the Acoustical Society of America*, 22(6):725–730, 1950.

[51] Joseph B Bayer, Nicole B Ellison, Sarita Y Schoenebeck, and Emily B Falk. Sharing the small moments: ephemeral social interaction on snapchat. *Information, Communication & Society*, 19(7):956–977, 2016.

[52] Ginestra Bianconi. The entropy of randomized network ensembles. *EPL (Europhysics Letters)*, 81(2):28005, 2007.

[53] Yvonne M Bishop, Stephen E Fienberg, and Paul W Holland. *Discrete multivariate analysis: theory and practice*. Springer Science & Business Media, 2007.

[54] Vincent D Blondel, Adeline Decuyper, and Gautier Krings. A survey of results on mobile phone datasets analysis. *EPJ Data Science*, 4(1):10, 2015.

[55] Stefano Boccaletti, Ginestra Bianconi, Regino Criado, Charo I Del Genio, Jesús Gómez-Gardenes, Miguel Romance, Irene Sendina-Nadal, Zhen Wang, and Massimiliano Zanin. The structure and dynamics of multilayer networks. *Physics Reports*, 544(1):1–122, 2014.

[56] Eszter Bokányi, Sándor Juhász, Márton Karsai, and Balázs Lengyel. Universal patterns of long-distance commuting and social assortativity in cities. *Scientific reports*, 11(1):1–10, 2021.

[57] Eszter Bokányi, Zsolt Vizi, Júlia Koltai, Gergely Röst, and Márton Karsai. Real-time estimation of the effective reproduction number of covid-19 from behavioral data. *arXiv preprint arXiv:2207.10738*, 2022.

[58] B. Bollobás and P. Erdős. *Graph theory and combinatorics: proceedings of the Cambridge combinatorial conference in honour of Paul Erdös*. Academic Press, Cambridge, 1984.

[59] S. P. Borgatti. Centrality and network flow. *Soc Netw*, 27, 2005.

[60] Javier Borge-Holthoefer, Raquel A Baños, Sandra González-Bailón, and Yamir Moreno. Cascading behaviour in complex socio-technical networks. *Journal of Complex Networks*, 1(1):3–24, 2013.

[61] Javier Borge-Holthoefer, Alejandro Rivero, Iñigo García, Elisa Cauhé, Alfredo Ferrer, Darío Ferrer, David Francos, David Iniguez, María Pilar Pérez, Gonzalo Ruiz, et al. Structural and dynamical patterns on online social networks: the spanish may 15th movement as a case study. *PloS one*, 6(8):e23883, 2011.

[62] Jean-Philippe Bouchaud. Crises and collective socio-economic phenomena: Simple models and challenges. *Journal of Statistical Physics*, 151(3-4):567–606, 2013.

[63] George E. P. Box, Gwilym M. Jenkins, and Gregory C. Reinsel. *Time series analysis : forecasting and control*. John Wiley, 4 edition, June 2008.

[64] Denis Boyer, Margaret C. Crofoot, and Peter D. Walsh. Non-random walks in monkeys and humans. *Journal of The Royal Society Interface*, 9(70):842–847, May 2012.

[65] U. Brandes. A faster algorithm for betweenness centrality. *J Math Sociol*, 25, 2001.

[66] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Comput Netw ISDN Syst*, 30, 1998.

[67] Charles D Brummitt, Raissa M D'Souza, and Elizabeth A Leicht. Suppressing cascades of load in interdependent networks. *Proceedings of the National Academy of Sciences*, 109(12):E680–E689, 2012.

[68] Charles D Brummitt, Kyu-Min Lee, and K-I Goh. Multiplexity-facilitated cascades in networks. *Physical Review E*, 85(4):045102, 2012.

[69] Michael Buhrmester, Tracy Kwang, and Samuel D Gosling. Amazon's mechanical turk: A new source of inexpensive, yet high-quality, data? *Perspectives on psychological science*, 6(1):3–5, 2011.

[70] Rebekka Burkholz, Antonios Garas, and Frank Schweitzer. How damage diversification can reduce systemic risk. *Physical Review E*, 93(4):042313, 2016.

[71] Jerome R Busemeyer and James T Townsend. Decision field theory: a dynamic-cognitive approach to decision making in an uncertain environment. *Psychological review*, 100(3):432, 1993.

[72] Daniel S Calovi, Ugo Lopez, Paul Schuhmacher, Hugues Chaté, Clément Sire, and Guy Theraulaz. Collective response to perturbations in a data-driven fish school model. *Journal of The Royal Society Interface*, 12(104):20141362, 2015.

[73] Claudio Castellano, Santo Fortunato, and Vittorio Loreto. Statistical physics of social dynamics. *Reviews of modern physics*, 81(2):591, 2009.

[74] Davide Cellai, Eduardo López, Jie Zhou, James P Gleeson, and Ginestra Bianconi. Percolation in multiplex networks with overlap. *Physical Review E*, 88(5):052811, 2013.

[75] Giulia Cencetti, Federico Battiston, Bruno Lepri, and Márton Karsai. Temporal properties of higher-order interactions in social networks. *Scientific reports*, 11(1):1–10, 2021.

[76] Damon Centola. The spread of behavior in an online social network experiment. *science*, 329(5996):1194–1197, 2010.

[77] Damon Centola. An experimental study of homophily in the adoption of health behavior. *Science*, 334(6060):1269–1272, 2011.

[78] Damon Centola and Michael Macy. Complex contagions and the weakness of long ties. *American journal of Sociology*, 113(3):702–734, 2007.

[79] Min Chen, Shiwen Mao, and Yunhao Liu. Big data: A survey. *Mobile Networks and Applications*, 19(2):171–209, 2014.

[80] E. Chlebus and W. Ludwin. Is handoff traffic really Poissonian? In *Proceedings of ICUPC '95 - 4th IEEE International Conference on Universal Personal Communications*, pages 348–353. IEEE, 1995.

[81] Robert B Cialdini and Noah J Goldstein. Social influence: Compliance and conformity. *Annu. Rev. Psychol.*, 55:591–621, 2004.

[82] Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman. Power-law distributions in empirical data. *SIAM review*, 51(4):661–703, 2009.

[83] Edith Cohen. Size-estimation framework with applications to transitive closure and reachability. *Journal of Computer and System Sciences*, 55(3):441–453, 1997.

[84] Reuven Cohen, Daniel Ben-Avraham, and Shlomo Havlin. Percolation critical exponents in scale-free networks. *Physical Review E*, 66(3):036113, 2002.

[85] Reuven Cohen and Shlomo Havlin. *Complex networks: structure, robustness and function*. Cambridge university press, 2010.

[86] Reuven Cohen, Shlomo Havlin, and Daniel Ben-Avraham. Efficient immunization strategies for computer networks and populations. *Physical review letters*, 91(24):247901, 2003.

[87] Brian Coley, Brigitte M. Jolles, Alain Farron, and Kamiar Aminian. Arm position during daily activity. *Gait & posture*, 28(4):581–587, November 2008.

[88] Vittoria Colizza, Romualdo Pastor-Satorras, and Alessandro Vespignani. Reaction–diffusion processes and metapopulation models in heterogeneous networks. *Nature Physics*, 3(4):276–282, 2007.

[89] Vittoria Colizza and Alessandro Vespignani. Invasion threshold in heterogeneous metapopulation networks. *Physical review letters*, 99(14):148701, 2007.

[90] Vittoria Colizza and Alessandro Vespignani. Epidemic modeling in metapopulation systems with heterogeneous coupling pattern: Theory and simulations. *Journal of theoretical biology*, 251(3):450–467, 2008.

[91] Thomas H Cormen, Charles E Leiserson, Ronald L Rivest, and Clifford Stein. *Introduction to algorithms*. MIT press, 2009.

[92] Álvaro Corral. Long-term clustering, scaling, and universality in the temporal occurrence of earthquakes. *Physical Review Letters*, 92:108501+, March 2004.

[93] Thomas M Cover and Joy A Thomas. Elements of information theory 2nd edition. *Willey-Interscience: NJ*, 2006.

[94] Scott Cox, KJ Horadam, and Asha Rao. The spread of ideas in a weighted threshold network. In *International Workshop on Complex Networks and their Applications*, pages 437–447. Springer, 2016.

[95] Pascal Crepey, Fabián P Alvarez, and Marc Barthélemy. Epidemic variability in complex networks. *Physical Review E*, 73(4):046131, 2006.

[96] Sicheng Dai, Hélène Bouchet, Márton Karsai, Eric Fleury, Jean-Pierre Chevrot, and Aurélie Nardy. Longitudinal data collection to follow social network and language development dynamics at preschool. *(Submitted)*, page 19, 2022.

[97] Sicheng Dai, Hélène Bouchet, Aurélie Nardy, Eric Fleury, Jean-Pierre Chevrot, and Márton Karsai. Temporal social network reconstruction using wireless proximity sensors: model selection and consequences. *EPJ Data Science*, 9(1):19, 2020.

[98] Daryl J Daley and Joe Gani. *Epidemic modelling: an introduction*, volume 15. Cambridge University Press, 2001.

[99] Daryl J Daley and David G Kendall. Epidemics and rumours. *Nature*, 204(4963):1118, 1964.

[100] Jörn Davidsen and Grzegorz Kwiatek. Earthquake interevent time distribution for induced micro-, nano-, and picoseismicity. *Physical Review Letters*, 110:068501+, February 2013.

[101] L. de Arcangelis, C. Godano, E. Lippiello, and M. Nicodemi. Universality in solar flare and earthquake occurrence. *Physical Review Letters*, 96(5):051102+, February 2006.

[102] Yves-Alexandre De Montjoye, César A Hidalgo, Michel Verleysen, and Vincent D Blondel. Unique in the crowd: The privacy bounds of human mobility. *Scientific reports*, 3:1376, 2013.

[103] Jaqueline Faria de Oliveira, Humberto Torres Marques-Neto, and Márton Karsai. Information adoption via repeated or diversified social influence on twitter. In *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 237–241. IEEE, 2020.

[104] Jaqueline Faria de Oliveira, Humberto Torres Marques-Neto, and Márton Karsai. Measuring the effects of repeated and diversified influence mechanism for information adoption on twitter. *Social Network Analysis and Mining*, 12(1):1–15, 2022.

[105] Z. Dezsö, E. Almaas, A. Lukács, B. Rácz, I. Szakadát, and A. L. Barabási. Dynamics of information access on the web. *Physical Review E*, 73(6):066132+, June 2006.

[106] Peter Sheridan Dodds, Kameron Decker Harris, and Christopher M Danforth. Limited imitation contagion on random networks: Chaos, universality, and unpredictability. *Physical review letters*, 110(15):158701, 2013.

[107] Peter Sheridan Dodds and Duncan J Watts. Universal behavior in a generalized model of contagion. *Physical review letters*, 92(21):218701, 2004.

[108] Sergey N Dorogovtsev and Jose FF Mendes. Evolution of networks. *Advances in physics*, 51(4):1079–1187, 2002.

[109] P Alex Dow, Lada A Adamic, and Adrien Friggeri. The anatomy of large facebook cascades. *ICWSM*, 1(2):12, 2013.

[110] RI Dunbar. The social brain hypothesis. *brain*, 9(10):178–190, 1998.

[111] Robin IM Dunbar. Cognitive constraints on the structure and dynamics of social networks. *Group Dynamics: Theory, Research, and Practice*, 12(1):7, 2008.

[112] Robin IM Dunbar. Social cognition on the internet: testing constraints on social network size. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1599):2192–2201, 2012.

[113] Nathan Eagle and Alex Sandy Pentland. Reality mining: sensing complex social systems. *Personal and ubiquitous computing*, 10(4):255–268, 2006.

[114] Ken TD Eames. Modelling disease spread through random and regular contacts in clustered populations. *Theoretical population biology*, 73(1):104–111, 2008.

[115] David Easley and Jon Kleinberg. *Networks, Crowds, and Markets: Reasoning about a Highly Connected World*. Cambridge University Press, Cambridge, 2010.

[116] Jean-Pierre Eckmann, Elisha Moses, and Danilo Sergi. Entropy of dialogues creates coherent structures in e-mail traffic. *Proceedings of the National Academy of Sciences of the United States of America*, 101(40):14333–14337, 2004.

[117] Christopher J Ellis and John Fender. Information cascades and revolutionary regime transitions. *The Economic Journal*, 121(553):763–792, 2011.

[118] Yrjö Engeström, Ritva Engeström, and Tarja Vähäaho. *When the center does not hold: The importance of knotworking*, volume 381. 1999.

[119] M. Ercsey-Ravasz and Z. Toroczkai. Centrality scaling in large networks. *Phys Rev Lett*, 105, 2010.

[120] Paul Erdős and A. Rényi. On the evolution of random graphs. *Publ. Math. Inst. Hungar. Acad. Sci*, 5:17–61, 1960.

[121] E. Estrada and J. A. Rodriguez-Velazquez. Subgraph centrality in complex networks. *Phys Rev E*, 71, 2005.

[122] M. G. Everett and S. P. Borgatti. Induced, endogenous and exogenous centrality. *Soc Netw*, 32, 2010.

[123] Joseph R Fauver, Mary E Petrone, Emma B Hodcroft, Kayoko Shioda, Hanna Y Ehrlich, Alexander G Watts, Chantal BF Vogels, Anderson F Brito, Tara Alpert, Anthony Muyombwe, et al. Coast-to-coast spread of sars-cov-2 during the early epidemic in the united states. *Cell*, 181(5):990–996, 2020.

[124] Juan Fernández-Gracia, Víctor M Eguíluz, and Maxi San Miguel. Timing interactions in social simulations: The voter model. In *Temporal Networks*, pages 331–352. Springer, 2013.

[125] Silvio C Ferreira, Ronan S Ferreira, and Romualdo Pastor-Satorras. Quasistationary analysis of the contact process on annealed scale-free networks. *Physical Review E*, 83(6):066113, 2011.

[126] Santo Fortunato. Community detection in graphs. *Physics reports*, 486(3):75–174, 2010.

[127] Ian Foster, Rayid Ghani, Ron S Jarmin, Frauke Kreuter, and Julia Lane. *Big data and social science: A practical guide to methods and tools*. CRC Press, 2016.

[128] James H Fowler and Nicholas A Christakis. Cooperative behavior cascades in human social networks. *Proceedings of the National Academy of Sciences*, page 200913149, 2010.

[129] Linton C Freeman. A set of measures of centrality based on betweenness. *Sociometry*, pages 35–41, 1977.

[130] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.

[131] Sabrina Gaito, Matteo Zignani, Gian Paolo Rossi, Alessandra Sala, Xiaohan Zhao, Haitao Zheng, and Ben Y Zhao. On the bursty evolution of online social networks. In *Proceedings of the First ACM International Workshop on Hot Topics on Interdisciplinary Social Networks Research*, pages 1–8. ACM, 2012.

[132] Jacques Gautrais, Christian Jost, Marc Soria, Alexandre Campo, Sébastien Motsch, Richard Fournier, Stéphane Blanco, and Guy Theraulaz. Analyzing fish movement as a persistent turning walker. *Journal of mathematical biology*, 58(3):429–445, 2009.

[133] Aurélien Gautreau, Alain Barrat, and Marc Barthélemy. Arrival time statistics in global disease spread. *Journal of Statistical Mechanics: Theory and Experiment*, 2007(09):L09001, 2007.

[134] Laetitia Gauvin, Mathieu Génois, Márton Karsai, Mikko Kivelä, Taro Takaguchi, Eugenio Valdano, and Christian L Vestergaard. Randomized reference models for temporal networks. *arXiv preprint arXiv:1806.04032*, 2018.

[135] Rumi Ghosh and Kristina Lerman. A framework for quantitative analysis of cascades on networks. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 665–674. ACM, 2011.

[136] Nigel Gilbert and Klaus Troitzsch. *Simulation for the social scientist*. McGraw-Hill Education (UK), 2005.

[137] Serena Giurgola, Simone Piaggesi, Márton Karsai, Yelena Mejova, André Panisson, and Michele Tizzoni. Mapping urban socioeconomic inequalities in developing countries through facebook advertising data. *arXiv preprint arXiv:2105.13774*, 2021.

[138] James P Gleeson. Cascades on correlated and modular random networks. *Physical Review E*, 77(4):046117, 2008.

[139] James P Gleeson. High-accuracy approximation of binary-state dynamics on networks. *Physical Review Letters*, 107(6):068701, 2011.

[140] James P Gleeson. Binary-state dynamics on complex networks: Pair approximation and beyond. *Physical Review X*, 3(2):021004, 2013.

[141] James P Gleeson and Diarmuid J Cahalane. Seed size strongly affects cascades on random networks. *Physical Review E*, 75(5):056103, 2007.

[142] James P Gleeson and Rick Durrett. Temporal profiles of avalanches on networks. *Nature communications*, 8(1):1–13, 2017.

[143] Sharad Goel, Duncan J Watts, and Daniel G Goldstein. The structure of online diffusion networks. In *Proceedings of the 13th ACM conference on electronic commerce*, pages 623–638. ACM, 2012.

[144] Michaela Goetz, Jure Leskovec, Mary McGlohon, and Christos Faloutsos. Modeling blog dynamics. In *International Conference on Weblogs and Social Media*, May 2009.

[145] W. Goffman and V. A. Newill. Generalization of epidemic theory: An application to the transmission of ideas. *Nature*, 204:225–228, 1964.

[146] Kwang-Il Goh and Albert-László Barabási. Burstiness and memory in complex systems. *EPL (Europhysics Letters)*, 81:48002+, February 2008.

[147] Bruno Gonçalves, Nicola Perra, and Alessandro Vespignani. Modeling users' activity on twitter networks: Validation of dunbar's number. *PloS one*, 6(8):e22656, 2011.

[148] Marta C Gonzalez, Cesar A Hidalgo, and Albert-László Barabási. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008.

[149] Sandra González-Bailón, Javier Borge-Holthoefer, Alejandro Rivero, and Yamir Moreno. The dynamics of protest recruitment through an online network. *Scientific reports*, 1:197, 2011.

[150] Palash Goyal and Emilio Ferrara. Graph embedding techniques. *Applications, and Performance: A Survey. arXiv preprint*, 2017.

[151] Clara Granell, Sergio Gómez, and Alex Arenas. Dynamical interplay between awareness and epidemic spreading in multiplex networks. *Physical review letters*, 111(12):128701, 2013.

[152] Mark Granovetter. Threshold models of collective behavior. *American journal of sociology*, 83(6):1420–1443, 1978.

[153] Mark Granovetter. The strength of weak ties: A network theory revisited. *Sociological theory*, pages 201–233, 1983.

[154] Mark Granovetter and Roland Soong. Threshold models of diffusion and collective behavior. *J. Math. Sociol*, 9(3):165–179, 1983.

[155] Mark S Granovetter. The strength of weak ties. *American journal of sociology*, 78(6):1360–1380, 1973.

[156] Geoffrey R. Grimmett and David Stirzaker. *Probability and random processes*. Oxford University Press, 3 edition, August 2009.

[157] Daniel Gruhl, Ramanathan Guha, David Liben-Nowell, and Andrew Tomkins. Information diffusion through blogspace. In *Proceedings of the 13th international conference on World Wide Web*, pages 491–501. ACM, 2004.

[158] Heidi Ann Hahn. The conundrum of verification and validation of social science-based models. *Procedia Computer Science*, 16:878–887, 2013.

[159] Henry E Hale. Regime change cascades: What we have learned from the 1848 revolutions to the 2011 arab uprisings. *Annual Review of Political Science*, 16:331–353, 2013.

[160] Xiao-Pu Han, Tao Zhou, and Bing-Hong Wang. Modeling human dynamics with adaptive interest. *New Journal of Physics*, 10(7):073010+, July 2008.

[161] U. Harder and M. Paczuski. Correlated dynamics in human printing behavior. *Physica A: Statistical Mechanics and its Applications*, 361(1):329–336, February 2006.

[162] Laurent Hébert-Dufresne, Edward Laurence, Antoine Allard, Jean-Gabriel Young, and Louis J Dubé. Complex networks as an emerging property of hierarchical preferential attachment. *Physical Review E*, 92(6):062809, 2015.

[163] Peter Hedström and Petri Ylikoski. Causal mechanisms in the social sciences. *Annual review of sociology*, 36:49–67, 2010.

[164] Fritz Heider. *The psychology of interpersonal relations*. Psychology Press, 2013.

[165] Stefan Heule, Marc Nunkesser, and Alexander Hall. Hyperloglog in practice: Algorithmic engineering of a state of the art cardinality estimation algorithm. In *Proceedings of the 16th International Conference on Extending Database Technology*, pages 683–692, 2013.

[166] Alison L Hill, David G Rand, Martin A Nowak, and Nicholas A Christakis. Infectious disease modeling of social contagion in networks. *PLOS computational biology*, 6(11):e1000968, 2010.

[167] Rafiazka Millanida Hilman, Gerardo Iñiguez, and Márton Karsai. Socioeconomic biases in urban mixing patterns of us metropolitan areas. *EPJ data science*, 11(1):32, 2022.

[168] Haye Hinrichsen. Non-equilibrium critical phenomena and phase transitions into absorbing states. *Advances in physics*, 49(7):815–958, 2000.

[169] Petter Holme. Network reachability of real-world contact sequences. *Physical Review E*, 71(4):046119, 2005.

[170] Petter Holme. Modern temporal network theory: a colloquium. *The European Physical Journal B*, 88(9):234, 2015.

[171] Petter Holme and Fredrik Liljeros. Birth and death of links control disease spreading in empirical contact networks. *Scientific reports*, 4:4999, 2014.

[172] Petter Holme and Fredrik Liljeros. Mechanistic models in computational social science. *Frontiers in Physics*, 3:78, 2015.

[173] Petter Holme and Jari Saramäki. Temporal networks. *Physics reports*, 519(3):97–125, 2012.

[174] Hadrien Hours, Eric Fleury, and Márton Karsai. Link prediction in the twitter mention network: impacts of local structure and similarity of interest. In *Data Mining Workshops (ICDMW), 2016 IEEE 16th International Conference on*, pages 454–461. IEEE, 2016.

[175] Thomas R Hurd and James P Gleeson. On watts' cascade model with random link weights. *Journal of Complex Networks*, 1(1):25–43, 2013.

[176] José Luis Iribarren and Esteban Moro. Impact of human activity patterns on the dynamics of information diffusion. *Physical review letters*, 103(3):038702, 2009.

[177] Ramakrishnan Iyer, Vilas Menon, Michael Buice, Christof Koch, and Stefan Mihalas. The influence of synaptic weight distribution on neuronal population dynamics. *PLoS computational biology*, 9(10):e1003248, 2013.

[178] Matthew O. Jackson. *Social and Economic Networks*. Princeton University Press, Princeton, 2008.

[179] Toni Janevski. *Traffic Analysis and Design of Wireless IP Networks*. Artech House, Inc., Norwood, MA, USA, 2003.

[180] Pablo Jensen, Matteo Morini, Márton Karsai, Tommaso Venturini, Alessandro Vespignani, Mathieu Jacomy, Jean-Philippe Cointet, Pierre Mercklé, and Eric Fleury. Detecting global bridges in networks. *Journal of Complex Networks*, 4(3):319–329, 2015.

[181] Hang-Hyun Jo, Márton Karsai, Juuso Karikoski, and Kimmo Kaski. Spatiotemporal correlations of handset-based service usages. *EPJ Data Science*, 1(1):10, 2012.

[182] Hang-Hyun Jo, Márton Karsai, János Kertész, and Kimmo Kaski. Circadian pattern and burstiness in mobile phone communication. *New Journal of Physics*, 14(1):013055, 2012.

[183] Hang-Hyun Jo, Yohsuke Murase, János Török, János Kertész, and Kimmo Kaski. Stylized facts in social networks: Community-based static modeling. *Physica A: Statistical Mechanics and its Applications*, 500:23–39, 2018.

[184] Hang-Hyun Jo, Juan I. Perotti, Kimmo Kaski, and János Kertész. Correlated bursts and the role of memory range. *Physical Review E*, 92(2):022814+, August 2015.

[185] Jeffrey Johnson. *Hypernetworks in the science of complex systems*. Imperial College Press, 2011.

[186] Tommi Junttila and Petteri Kaski. Engineering an efficient canonical labeling tool for large and sparse graphs. In *2007 Proceedings of the Ninth Workshop on Algorithm Engineering and Experiments (ALENEX)*, pages 135–149. SIAM, 2007.

[187] Eirini Kalliamvakou, Georgios Gousios, Kelly Blincoe, Leif Singer, Daniel M German, and Daniela Damian. The promises and perils of mining github. In *Proceedings of the 11th working conference on mining software repositories*, pages 92–101. ACM, 2014.

[188] Dayun Kang, Hyunho Choi, Jong-Hun Kim, and Jungsoon Choi. Spatial epidemic dynamics of the covid-19 outbreak in china. *International Journal of Infectious Diseases*, 94:96–102, 2020.

[189] Panagiotis D Karampourniotis, Sameet Sreenivasan, Boleslaw K Szymanski, and Gyorgy Korniss. The impact of heterogeneous thresholds on social contagion with multiple initiators. *PloS one*, 10(11):e0143020, 2015.

[190] Fariba Karimi and Petter Holme. A temporal network version of watts's cascade model. In *Temporal Networks*, pages 315–329. Springer, 2013.

[191] Fariba Karimi and Petter Holme. Threshold model of cascades in empirical temporal networks. *Physica A: Statistical Mechanics and its Applications*, 392(16):3476–3483, 2013.

[192] Fariba Karimi and Petter Holme. Threshold model of cascades in empirical temporal networks. *Physica A: Statistical Mechanics and its Applications*, 392(16):3476–3483, 2013.

[193] M Karsai, J-Ch Anglès d'Auriac, and F Iglói. Rounding of first-order phase transitions and optimal cooperation in scale-free networks. *Physical Review E*, 76(4):041107, 2007.

[194] M Karsai, J-Ch Anglès d'Auriac, and F Iglói. Nonequilibrium dynamics of fully frustrated ising models at t= 0. *Journal of Statistical Mechanics: Theory and Experiment*, 2009(07):P07044, 2009.

[195] M Karsai, M Kivelä, RK Pan, K Kaski, J Kertész, A-L Barabási, and J Saramäki. Small but slow world: How network topology and burstiness slow down spreading. *Physical Review E*, 83(2):025102, 2011.

[196] M Karsai, IA Kovács, J-Ch Anglès d'Auriac, and F Iglói. Density of critical clusters in strips of strongly disordered systems. *Physical Review E*, 78(6):061109, 2008.

[197] Márton Karsai, J-Ch Anglès d'Auriac, and F Iglói. Interface mapping in two-dimensional random lattice models. *Journal of Statistical Mechanics: Theory and Experiment*, 2010(08):P08027, 2010.

[198] Márton Karsai, Gerardo Iniguez, Kimmo Kaski, and János Kertész. Complex contagion process in spreading of online innovation. *Journal of The Royal Society Interface*, 11(101):20140694, 2014.

[199] Márton Karsai, Gerardo Iñiguez, Riivo Kikas, Kimmo Kaski, and János Kertész. Local cascades induced global contagion: How heterogeneous thresholds, exogenous effects, and unconcerned behaviour govern online adoption spreading. *Scientific reports*, 6:27178, 2016.

[200] Márton Karsai, Hang-Hyun Jo, and Kimmo Kaski. *Bursty human dynamics*. Springer, 2018.

[201] Márton Karsai, Róbert Juhász, and Ferenc Iglói. Nonequilibrium phase transitions and finite-size scaling in weighted scale-free networks. *Physical Review E*, 73(3):036116, 2006.

[202] Márton Karsai, Kimmo Kaski, Albert-László Barabási, and János Kertész. Universal features of correlated bursty behaviour. *Scientific reports*, 2:397, 2011.

[203] Márton Karsai, Kimmo Kaski, and János Kertész. Correlated dynamics in egocentric communication networks. *Plos one*, 7(7):e40612, 2012.

[204] Márton Karsai, Júlia Koltai, Orsolya Vásárhelyi, and Gergely Röst. Hungary in mask/maszk in hungary. *Corvinus Journal of Sociology and Social Policy*, (2), 2020.

[205] Márton Karsai and Nicola Perra. Control strategies of contagion processes in time-varying networks. In *Temporal Network Epidemiology*, pages 179–197. Springer, Singapore, 2017.

[206] Márton Karsai, Nicola Perra, and Alessandro Vespignani. Time varying networks and the weakness of strong ties. *Scientific reports*, 4:4001, 2014.

[207] Leo Katz. A new status index derived from sociometric analysis. *Psychometrika*, 18(1):39–43, 1953.

[208] Ichiro Kawachi and Lisa F Berkman. Social ties and mental health. *Journal of Urban health*, 78(3):458–467, 2001.

[209] Matt J Keeling and Pejman Rohani. *Modeling infectious diseases in humans and animals*. Princeton University Press, 2011.

[210] David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 137–146. ACM, 2003.

[211] Takehito Kemuriyama, Hiroyuki Ohta, Yoshiaki Sato, Satoshi Maruyama, Megumi Tandai-Hiruma, Kazuo Kato, and Yasuhiro Nishida. A power-law distribution of inter-spike intervals in renal sympathetic nerve activity in salt-sensitive hypertension-induced chronic heart failure. *BioSystems*, 101(2):144–147, 2010.

[212] Riivo Kikas, Marlon Dumas, and Márton Karsai. Bursty egocentric network evolution in skype. *Social Network Analysis and Mining*, 3(4):1393–1401, 2013.

[213] Eun-Kyeong Kim and Hang-Hyun Jo. Measuring burstiness for finite event sequences. *Physical Review E*, 94:032311+, September 2016.

[214] Maksim Kitsak, Lazaros Gallos, Shlomo Havlin, Fredrik Liljeros, Lev Muchnik, H. Stanley, and Hernan Makse. Identification of influential spreaders in complex networks. *Nature Physics*, 6, 01 2010.

[215] Maksim Kitsak, Lazaros K Gallos, Shlomo Havlin, Fredrik Liljeros, Lev Muchnik, H Eugene Stanley, and Hernán A Makse. Identification of influential spreaders in complex networks. *Nature physics*, 6(11):888, 2010.

[216] Mikko Kivelä, Alex Arenas, Marc Barthelemy, James P Gleeson, Yamir Moreno, and Mason A Porter. Multilayer networks. *Journal of complex networks*, 2(3):203–271, 2014.

[217] Mikko Kivelä, Jordan Cambe, Jari Saramäki, and Márton Karsai. Mapping temporal-network percolation to weighted, static event graphs. *Scientific reports*, 8(1):12357, 2018.

[218] Mikko Kivelä, Raj Kumar Pan, Kimmo Kaski, János Kertész, Jari Saramäki, and Márton Karsai. Multiscale analysis of spreading in a large communication network. *Journal of Statistical Mechanics: Theory and Experiment*, page P03005, 2012.

[219] Stephen D. Kleban and Scott H. Clearwater. Hierarchical dynamics, interarrival times, and performance. In *Proceedings of the 2003 ACM/IEEE Conference on Supercomputing*, SC '03, New York, NY, USA, 2003. ACM.

[220] Peter Klimek, Renaud Lambiotte, and Stefan Thurner. Opinion formation in laggard societies. *EPL (Europhysics Letters)*, 82(2):28008, 2008.

[221] Júlia Koltai, Orsolya Vásárhelyi, Gergely Röst, and Márton Karsai. Reconstructing social mixing patterns via weighted contact matrices from online and representative surveys. *Scientific Reports*, 12(1):1–12, 2022.

[222] Gueorgi Kossinets and Duncan J Watts. Empirical analysis of an evolving social network. *Science*, 311(5757):88–90, 2006.

[223] Lauri Kovanen, Márton Karsai, Kimmo Kaski, János Kertész, and Jari Saramäki. Temporal motifs in time-dependent networks. *Journal of Statistical Mechanics: Theory and Experiment*, page P11005, 2011.

[224] Lauri Kovanen, Márton Karsai, Kimmo Kaski, János Kertész, and Jari Saramäki. Temporal motifs. In *Temporal Networks*, pages 119–133. Springer, Berlin, Heidelberg, 2013.

[225] Moritz UG Kraemer, Chia-Hung Yang, Bernardo Gutierrez, Chieh-Hsi Wu, Brennan Klein, David M Pigott, Louis Du Plessis, Nuno R Faria, Ruoran Li, William P Hanage, et al. The effect of human mobility and control measures on the covid-19 epidemic in china. *Science*, 368(6490):493–497, 2020.

[226] Gautier Krings, Márton Karsai, Sebastian Bernhardsson, Vincent D Blondel, and Jari Saramäki. Effects of time window size and placement on the structure of an aggregated communication network. *EPJ Data Science*, 1(1):4, 2012.

[227] Janos M Kumpula, J-P Onnela, Jari Saramäki, Janos Kertész, and Kimmo Kaski. Model of community emergence in weighted social networks. *Computer Physics Communications*, 180(4):517–522, 2009.

[228] Jussi M Kumpula, Jukka-Pekka Onnela, Jari Saramäki, Kimmo Kaski, and János Kertész. Emergence of communities in weighted networks. *Physical review letters*, 99(22):228701, 2007.

[229] Kristóf Kutasi, Júlia Koltai, Ágnes Szabó-Morvai, Gergely R ost, Márton Karsai, Péter Biró, and Balázs Lengyel. Time varying networks and the weakness of strong ties. *Scientific reports*, 12:13293, 2022.

[230] Renaud Lambiotte, Martin Rosvall, and Ingo Scholtes. Understanding complex systems: From networks to optimal higher-order models. *arXiv preprint arXiv:1806.05977*, 2018.

[231] Bibb Latané. The psychology of social impact. *American psychologist*, 36(4):343, 1981.

[232] Matthieu Latapy, Tiphaine Viard, and Clémence Magnien. Stream graphs and link streams for the modeling of interactions over time. *Social Network Analysis and Mining*, 8(1):1–29, 2018.

[233] Guillaume Laurent, Jari Saramäki, and Márton Karsai. From calls to communities: a model for time-varying social networks. *The European Physical Journal B*, 88(11):301, 2015.

[234] David Lazer, Alex Sandy Pentland, Lada Adamic, Sinan Aral, Albert Laszlo Barabási, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, et al. Life in the network: the coming age of computational social science. *Science (New York, NY)*, 323(5915):721, 2009.

[235] Kyu-Min Lee, Charles D Brummitt, and K-I Goh. Threshold cascades with response heterogeneity in multiplex networks. *Physical Review E*, 90(6):062816, 2014.

[236] Kyunghan Lee, Joohyun Lee, Yung Yi, Injong Rhee, and Song Chong. Mobile data offloading: How much can wifi deliver? *IEEE/ACM Transactions on Networking*, 21(2):536–550, April 2013.

[237] Yannick Leo, Eric Fleury, J Ignacio Alvarez-Hamelin, Carlos Sarraute, and Márton Karsai. Socioeconomic correlations and stratification in social-communication networks. *Journal of The Royal Society Interface*, 13(125):20160598, 2016.

[238] Yannick Leo, Márton Karsai, Carlos Sarraute, and Eric Fleury. Correlations of consumption patterns in social-economic networks. In *Advances in Social Networks Analysis and Mining (ASONAM), 2016 IEEE/ACM International Conference on*, pages 493–500. IEEE, 2016.

[239] Yannick Leo, Márton Karsai, Carlos Sarraute, and Eric Fleury. Correlations and dynamics of consumption patterns in social-economic networks. *Social Network Analysis and Mining*, 8:9, 2018.

[240] W. W. Leontief. *The structure of American economy, 1919–1929: an empirical application of equilibrium analysis*. Harvard University Press, Cambridge, 1941.

[241] Sébastien Lerique, Jacob Levy Abitbol, and Márton Karsai. Joint embedding of structure and features via graph convolutional networks. *Applied Network Science*, 5(1):1–24, 2020.

[242] Jure Leskovec, Lada A Adamic, and Bernardo A Huberman. The dynamics of viral marketing. *ACM Transactions on the Web (TWEB)*, 1(1):5, 2007.

[243] Jure Leskovec, Ajit Singh, and Jon Kleinberg. Patterns of influence in a recommendation network. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 380–389. Springer, 2006.

[244] Jacob Levy Abitbol, Eric Fleury, and Márton Karsai. Optimal proxy selection for socioeconomic status inference on twitter. *Complexity*, 2019, 2019.

[245] Jacobo Levy Abitbol, Márton Karsai, and Eric Fleury. Location, occupation, and semantics based socioeconomic status inference on twitter. In *ICDM'18*, 2018.

[246] Yongjun Liao, Wei Du, Márton Karsai, Carlos Sarraute, Martin Minnoni, and Eric Fleury. Prepaid or postpaid? that is the question: novel methods of subscription type prediction in mobile phone services. In *Social Network Based Big Data Analysis and Applications*, pages 165–181. Springer, 2018.

[247] Christian Licoppe and Zbigniew Smoreda. Are social networks technologically embedded?: How networks are changing today with changes in communication technology. *Social networks*, 27(4):317–335, 2005.

[248] Anna Sára Ligeti, Márton Karsai, Júlia Anna Koltai, and Gergely Röst. A covid-19 elleni védőoltással kapcsolatos attitűdök a magyar társadalomban. 2022.

[249] Jessica Lin, Eamonn Keogh, Stefano Lonardi, and Bill Chiu. A symbolic representation of time series, with implications for streaming algorithms. In *Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery*, pages 2–11. ACM, 2003.

[250] Suyu Liu, Nicola Perra, Márton Karsai, and Alessandro Vespignani. Controlling contagion processes in activity driven networks. *Physical review letters*, 112(11):118702, 2014.

[251] S. B. Lowen and M. C. Teich. Fractal renewal processes generate 1/f noise. *Phys. Rev. E*, 47:992–1001, February 1993.

[252] Jean-Christophe Lucet, Cédric Laouenan, Guillaume Chelius, Nicolas Veziris, Didier Lepelletier, Adrien Friggeri, Dominique Abiteboul, Elisabeth Bouvet, France Mentre, and Eric Fleury. Electronic sensors for assessing interactions between healthcare workers and patients under airborne precautions. *PloS one*, 7(5):e37893, 2012.

[253] Hosam Mahmoud. *Pólya urn models*. Chapman and Hall/CRC, 2008.

[254] R. Dean Malmgren, Daniel B. Stouffer, Andriana S. L. O. Campanharo, and Luis A. Amaral. On universality in human correspondence activity. *Science (New York, N.Y.)*, 325(5948):1696–1700, September 2009.

[255] R. Dean Malmgren, Daniel B. Stouffer, Adilson E. Motter, and Luís A. N. Amaral. A poissonian explanation for heavy tails in e-mail communication. *Proceedings of the National Academy of Sciences*, 105(47):18153–18158, November 2008.

[256] Rosario N. Mantegna and H. Eugene Stanley. *Introduction to Econophysics: Correlations and Complexity in Finance*. Cambridge University Press, 1 edition, July 2007.

[257] Naoki Masuda and Petter Holme. *Temporal Network Epidemiology*. Springer, 2017.

[258] Naoki Masuda and Renaud Lambiotte. *A Guide to Temporal Networks*. World Scientific, 2016.

[259] Naoki Masuda, Taro Takaguchi, Nobuo Sato, and Kazuo Yano. Self-Exciting Point Process Modeling of Conversation Event Sequences. In Petter Holme and Jari Saramäki, editors, *Temporal Networks*, Understanding Complex Systems, pages 245–264+. Springer Berlin Heidelberg, 2013.

[260] Viktor Mayer-Schönberger and Kenneth Cukier. Big data: A revolution that transforms how we work, live, and think, 2012.

[261] R. T. James McAteer, C. Alex Young, Jack Ireland, and Peter T. Gallagher. The Bursty Nature of Solar Flare X-Ray Emission. *The Astrophysical Journal*, 662(1):691, 2007.

[262] Miller McPherson, Lynn Smith-Lovin, and James M Cook. Birds of a feather: Homophily in social networks. *Annual review of sociology*, 27(1):415–444, 2001.

[263] Andrew Mellor. The temporal event graph. *Journal of Complex Networks*, 2017.

[264] Sergey Melnik, Jonathan A Ward, James P Gleeson, and Mason A Porter. Multi-stage complex contagions. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 23(1):013124, 2013.

[265] Ziad A Memish, Nawfal Aljerian, and Shahul H Ebrahim. Tale of three seeding patterns of sars-cov-2 in saudi arabia. *The Lancet Infectious Diseases*, 21(1):26–27, 2021.

[266] Ron Milo, Shai Shen-Orr, Shalev Itzkovitz, Nadav Kashtan, Dmitri Chklovskii, and Uri Alon. Network motifs: simple building blocks of complex networks. *Science*, 298(5594):824–827, 2002.

[267] Giovanna Miritello, Rubén Lara, Manuel Cebrian, and Esteban Moro. Limited communication capacity unveils strategies for human interaction. *Scientific reports*, 3:1950, 2013.

[268] Giovanna Miritello, Esteban Moro, and Rubén Lara. Dynamical strength of social ties in information spreading. *Physical Review E*, 83(4):045102, 2011.

[269] Delia Mocanu, Luca Rossi, Qian Zhang, Marton Karsai, and Walter Quattrociocchi. Collective attention in the age of (mis) information. *Computers in Human Behavior*, 51:1198–1204, 2015.

[270] Michael Molloy and Bruce Reed. A critical point for random graphs with a given degree sequence. *Random structures & algorithms*, 6(2-3):161–180, 1995.

[271] Martina Morris. Telling tails explain the discrepancy in sexual partner reports. *Nature*, 365(6445):437, 1993.

[272] R. C. Morrissey, N. D. Goldman, and K. P. Kennedy. Skype s.a. united states security registration statement, amendment 3, reg. no. 333-168646, 2011. Accessed: 2014.10.14.

[273] Peter J Mucha, Thomas Richardson, Kevin Macon, Mason A Porter, and Jukka-Pekka Onnela. Community structure in time-dependent, multiscale, and multiplex networks. *science*, 328(5980):876–878, 2010.

[274] Seth A. Myers and Jure Leskovec. The bursty dynamics of the twitter information network. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14, pages 913–924, New York, NY, USA, 2014. ACM.

[275] Toru Nakamura, Toru Takumi, Atsuko Takano, Naoko Aoyagi, Kazuhiro Yoshiuchi, Zbigniew R. Struzik, and Yoshiharu Yamamoto. Of mice and men — universality and breakdown of behavioral organization. *PLoS ONE*, 3(4):e2050+, April 2008.

[276] Neema Nassir, Mark Hickman, Ali Malekzadeh, and Elnaz Irannezhad. A utility-based travel impedance measure for public transit network accessibility. *Transportation Research Part A: Policy and Practice*, 88:26–39, 2016.

[277] M. E. J. Newman. Power laws, pareto distributions and zipf's law. *Contemporary Physics*, 46(5):323–351+, 2005.

[278] Mark Newman. *Networks: an introduction*. Oxford university press, 2010.

[279] Mark Newman. *Networks*. Oxford university press, 2018.

[280] Mark EJ Newman. The structure and function of complex networks. *SIAM review*, 45(2):167–256, 2003.

[281] Mark EJ Newman, Steven H Strogatz, and Duncan J Watts. Random graphs with arbitrary degree distributions and their applications. *Physical review E*, 64(2):026118, 2001.

[282] Eric WT Ngai, Spencer SC Tao, and Karen KL Moon. Social media research: Theories, constructs, and conceptual frameworks. *International Journal of Information Management*, 35(1):33–44, 2015.

[283] Vincenzo Nicosia, John Tang, Mirco Musolesi, Giovanni Russo, Cecilia Mascolo, and Vito Latora. Components in time-varying graphs. *Chaos: An interdisciplinary journal of nonlinear science*, 22(2):023101, 2012.

[284] Jae Dong Noh and Hyunggyu Park. Critical behavior of the contact process in annealed scale-free networks. *Physical Review E*, 79(5):056115, 2009.

[285] Gergely Ódor, Domonkos Czifra, Júlia Komjáthy, László Lovász, and Márton Karsai. Switchover phenomenon induced by epidemic seeding on geometric networks. *Proceedings of the National Academy of Sciences*, 118(41), 2021.

[286] Géza Odor. Slow, bursty dynamics as a consequence of quenched network topologies. *Physical Review E*, 89(4):042102, 2014.

[287] Joao G. Oliveira and Albert-Laszlo Barabási. Human dynamics: Darwin and einstein correspondence patterns. *Nature*, 437(7063):1251, October 2005.

[288] J-P Onnela, Jari Saramäki, Jorkki Hyvönen, György Szabó, David Lazer, Kimmo Kaski, János Kertész, and A-L Barabási. Structure and tie strengths in mobile communication networks. *Proceedings of the national academy of sciences*, 104(18):7332–7336, 2007.

[289] Jukka-Pekka Onnela and Felix Reed-Tsochas. Spontaneous emergence of social influence in online systems. *Proceedings of the National Academy of Sciences*, 107(43):18375–18380, 2010.

[290] R Lyman Ott and Micheal T Longnecker. *An introduction to statistical methods and data analysis*. Nelson Education, 2015.

[291] Vasyl Palchykov, Kimmo Kaski, Janos Kertész, Albert-László Barabási, and Robin IM Dunbar. Sex differences in intimate relationships. *Scientific reports*, 2:370, 2012.

[292] Raj Kumar Pan and Jari Saramäki. Path lengths, correlations, and centrality in temporal networks. *Physical Review E*, 84(1):016105, 2011.

[293] Romualdo Pastor-Satorras and Alessandro Vespignani. Epidemic spreading in scale-free networks. *Physical review letters*, 86(14):3200, 2001.

[294] Romualdo Pastor-Satorras and Alessandro Vespignani. Immunization of complex networks. *Physical Review E*, 65(3):036104, 2002.

[295] Romualdo Pastor-Satorras and Alessandro Vespignani. *Evolution and structure of the Internet: A statistical physics approach*. Cambridge University Press, 2007.

[296] Vern Paxson and Sally Floyd. Wide-area traffic: The failure of poisson modeling. *IEEE/ACM Transactions on Networking*, 3:226–244, 1995.

[297] Tiago P Peixoto and Martin Rosvall. Modelling sequences and temporal networks with dynamic community structures. *Nature communications*, 8(1):582, 2017.

[298] Tamás Péni, Balázs Csutak, Gábor Szederkényi, and Gergely Röst. Nonlinear model predictive control with logic constraints for covid-19 management. *Nonlinear Dynamics*, 102(4):1965–1986, 2020.

[299] Nicola Perra, Bruno Gonçalves, Romualdo Pastor-Satorras, and Alessandro Vespignani. Activity driven modeling of time varying networks. *Scientific reports*, 2:469, 2012.

[300] Giovanni Petri, Martina Scolamiero, Irene Donato, and Francesco Vaccarino. Topological strata of weighted complex networks. *PloS one*, 8(6):e66506, 2013.

[301] Michael Plischke. *Equilibrium statistical physics*. World Scientific, 2006.

[302] Mason A Porter and James P Gleeson. Dynamical systems on networks. *Frontiers in Applied Dynamical Systems: Reviews and Tutorials*, 4, 2016.

[303] B Aditya Prakash, Hanghang Tong, Nicholas Valler, Michalis Faloutsos, and Christos Faloutsos. Virus propagation on time-varying networks: Theory and immunization algorithms. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 99–114. Springer, 2010.

[304] Kiesha Prem, Alex R Cook, and Mark Jit. Projecting social contact matrices in 152 countries using contact surveys and demographic data. *PLoS Computational Biology*, 13(9):e1005697, 2017.

[305] Alex Proekt, Jayanth R Banavar, Amos Maritan, and Donald W Pfaff. Scale invariance in the dynamics of spontaneous behavior. *Proceedings of the National Academy of Sciences*, 109(26):10564–10569, 2012.

[306] Hassan Qudrat-Ullah, J Michael Spector, and Paal Davidsen. *Complex decision making: Theory and practice*. springer, 2007.

[307] Anatol Rapoport. Spread of information through a population with socio-structural bias: I. assumption of transitivity. *Bulletin of Mathematical Biology*, 15(4):523–533, 1953.

[308] Anatol Rapoport. Contribution to the theory of random and biased nets. In *Social Networks*, pages 389–409. Elsevier, 1977.

[309] Ulf-Dietrich Reips. The methodology of internet-based experiments. *The Oxford handbook of Internet psychology*, pages 373–390, 2007.

[310] Bruno Ribeiro, Nicola Perra, and Andrea Baronchelli. Quantifying the effect of temporal resolution on time-varying networks. *Scientific reports*, 3:3006, 2013.

[311] Luis EC Rocha, Fredrik Liljeros, and Petter Holme. Simulated epidemics in an empirical spatiotemporal network of 50,185 sexual contacts. *PLoS computational biology*, 7(3):e1001109, 2011.

[312] Ignacio Rodriguez-Iturbe and Andrea Rinaldo. *Fractal river basins: chance and self-organization*. Cambridge University Press, 2001.

[313] Everett M Rogers. *Diffusion of innovations*. Simon and Schuster, 2010.

[314] Daniel M Romero, Brendan Meeder, and Jon Kleinberg. Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. In *Proceedings of the 20th international conference on World wide web*, pages 695–704. ACM, 2011.

[315] Gergely Röst, Ferenc A Bartha, Norbert Bogya, Péter Boldog, Attila Dénes, Tamás Ferenci, Krisztina J Horváth, Attila Juhász, Csilla Nagy, Tamás Tekeli, et al. Early phase of the COVID-19 outbreak in Hungary and post-lockdown scenarios. *Viruses*, 12(7):708, 2020.

[316] Zhongyuan Ruan, Gerardo Iniguez, Márton Karsai, and János Kertész. Kinetics of social contagion. *Physical review letters*, 115(21):218702, 2015.

[317] Csilla Rudas, Olivér Surányi, Taha Yasseri, and János Török. Understanding and coping with extremism in an online collaborative environment: A data-driven modeling. *PloS one*, 12(3):e0173561, 2017.

[318] Jeffrey D Scargle. Studies in astronomical time series analysis. v. bayesian blocks, a new method to analyze structure in photon counting data. *The Astrophysical Journal*, 504(1):405, 1998.

[319] Thomas C Schelling. Models of segregation. *The American Economic Review*, 59(2):488–493, 1969.

[320] Ingo Scholtes, Nicolas Wider, René Pfitzner, Antonios Garas, Claudio J Tessone, and Frank Schweitzer. Causality-driven slow-down and speed-up of diffusion in non-markovian temporal networks. *Nature communications*, 5:5024, 2014.

[321] Esther Shein. Ephemeral data. *Communications of the ACM*, 56(9):20–22, 2013.

[322] Georg Simmel. 1908: Soziologie. *Untersuchungen über die Formen der Vergesellschaftung. Frankfurt am Main: Suhrkamp*, 1992.

[323] Georg Simmel and Kurt Heinrich WOLFF. *The Sociology of Georg Simmel. Translated, edited, and with an introduction by Kurt H. Wolff*. New York; Collier-Macmillan: London, 1964.

[324] Pramesh Singh, Sameet Sreenivasan, Boleslaw K Szymanski, and Gyorgy Korniss. Threshold-limited spreading in social networks with multiple initiators. *Scientific reports*, 3:2330, 2013.

[325] A. Skalka. *Understanding Human Dynamics: Report of the Defense Science Board Task Force*. DIANE Publishing Company, 2009.

[326] R. F. Smalley, J. L. Chatelain, D. L. Turcotte, and R. Prévot. A fractal approach to the clustering of earthquakes: Applications to the seismicity of the New Hebrides. *Bulletin of the Seismological Society of America*, 77(4):1368–1381, August 1987.

[327] Neil J Smelser. *Theory of collective behavior*. Quid Pro Books, 2011.

[328] Amanda Sorribes, Beatriz G Armendariz, Diego Lopez-Pigozzi, Cristina Murga, and Gonzalo G de Polavieja. The origin of behavioral bursts in decision-making circuitry. *PLoS computational biology*, 7(6):e1002075, 2011.

[329] H. Eugene Stanley. *Introduction to Phase Transitions and Critical Phenomena (International Series of Monographs on Physics)*. Oxford University Press, USA, July 1987.

[330] Michele Starnini and Romualdo Pastor-Satorras. Topological properties of a time-integrated activity-driven network. *Physical Review E*, 87(6):062807, 2013.

[331] Juliette Stehlé, Alain Barrat, and Ginestra Bianconi. Dynamical and bursty interactions in social networks. *Physical review E*, 81(3):035101, 2010.

[332] Juliette Stehlé, Nicolas Voirin, Alain Barrat, Ciro Cattuto, Vittoria Colizza, Lorenzo Isella, Corinne Régis, Jean-François Pinton, Nagham Khanafer, Wouter Van den Broeck, et al. Simulation of an seir infectious disease model on the dynamic contact network of conference attendees. *BMC medicine*, 9(1):87, 2011.

[333] James P Stevens. *Applied multivariate statistics for the social sciences*. Routledge, 2012.

[334] Arkadiusz Stopczynski, Vedran Sekara, Piotr Sapiezynski, Andrea Cuttone, Mette My Madsen, Jakob Eg Larsen, and Sune Lehmann. Measuring large-scale social networks with high resolution. *PloS one*, 9(4):e95978, 2014.

[335] Siddharth Suri and Duncan J Watts. Cooperation and contagion in web-based, networked public goods experiments. *PloS one*, 6(3):e16836, 2011.

[336] Miklos Szocska, Peter Pollner, Istvan Schiszler, Tamas Joo, Tamas Palicz, Martin McKee, Aron Asztalos, Laszlo Bencze, Mor Kapronczay, Peter Petrecz, et al. Countrywide population movement monitoring using mobile devices generated (big) data during the covid-19 crisis. *Scientific reports*, 11(1):1–9, 2021.

[337] Taro Takaguchi, Naoki Masuda, and Petter Holme. Bursty communication patterns facilitate spreading in a threshold-based epidemic dynamics. *PloS one*, 8(7):e68629, 2013.

[338] Da-Hai Tang, Xiao-Pu Han, and Bing-Hong Wang. Stretched exponential distribution of recurrent time of wars in china. *Physica A: Statistical Mechanics and its Applications*, 389(13):2637–2641, July 2010.

[339] Gergely Tibély, Lauri Kovanen, Márton Karsai, Kimmo Kaski, János Kertész, and Jari Saramäki. Communities and beyond: mesoscopic analysis of a large social network with complementary methods. *Physical Review E*, 83(5):056125, 2011.

[340] Michele Tizzoni, Paolo Bajardi, Chiara Poletto, José J Ramasco, Duygu Balcan, Bruno Gonçalves, Nicola Perra, Vittoria Colizza, and Alessandro Vespignani. Real-time numerical forecast of global epidemic spreading: case study of 2009 a/h1n1pdm. *BMC medicine*, 10(1):165, 2012.

[341] Michele Tizzoni, Elaine O Nsoesie, Laetitia Gauvin, Márton Karsai, Nicola Perra, and Shweta Bansal. Addressing the socioeconomic divide in computational modeling for infectious diseases. *Nature Communications*, 13(1):1–7, 2022.

[342] Michele Tizzoni, Kaiyuan Sun, Diego Benusiglio, Márton Karsai, and Nicola Perra. The scaling of human contacts and epidemic processes in metapopulation networks. *Scientific reports*, 5:15111, 2015.

[343] Riitta Toivonen, Xavier Castelló, Víctor M Eguíluz, Jari Saramäki, K Kaski, and Maxi San Miguel. Broad lifetime distributions for ordering dynamics in complex networks. *Physical Review E*, 79(1):016109, 2009.

[344] Mario V Tomasello, Nicola Perra, Claudio J Tessone, Márton Karsai, and Frank Schweitzer. The role of endogenous and exogenous mechanisms in the formation of r&d networks. *Scientific reports*, 4:5679, 2014.

[345] Jameson L Toole, Meeyoung Cha, and Marta C González. Modeling the adoption of innovations in the presence of geographic and media influences. *PloS one*, 7(1):e29528, 2012.

[346] Maddalena Torricelli, Márton Karsai, and Laetitia Gauvin. weg2vec: Event embedding for temporal networks. *Scientific Reports*, 10(1):1–11, 2020.

[347] B. A. N. Travençolo and L. d. F. Costa. Accessibility in complex networks. *Phys Lett A*, 373, 2008.

[348] Carolina Tripp-Barba, Cristina Alcaraz, and Mónica Aguilar Igartua. Special issue on modeling and performance evaluation of wireless ad-hoc networks. *Ad Hoc Networks*, 52(C):1–2, 2016.

[349] Lon Turnbull, Emese Dian, and Guenter Gross. The string method of burst identification in neuronal spike trains. *Journal of neuroscience methods*, 145(1-2):23–35, 2005.

[350] John C Turner. Social influence. thomson brooks, 1991.

[351] Enrico Ubaldi, Nicola Perra, Márton Karsai, Alessandro Vezzani, Raffaella Burioni, and Alessandro Vespignani. Asymptotic theory of time-varying social networks with heterogeneous activity and tie allocation. *Scientific reports*, 6:35724, 2016.

[352] Enrico Ubaldi, Alessandro Vezzani, Márton Karsai, Nicola Perra, and Raffaella Burioni. Burstiness and tie activation strategies in time-varying social networks. *Scientific reports*, 7:46225, 2017.

[353] Samuel Unicomb, Gerardo Iñiguez, James P Gleeson, and Márton Karsai. Dynamics of cascades on burstiness-controlled temporal networks. *Nature communications*, 12(1):1–10, 2021.

[354] Samuel Unicomb, Gerardo Iñiguez, and Márton Karsai. Threshold driven contagion on weighted networks. *Scientific reports*, 8(1):3094, 2018.

[355] Samuel Unicomb, Gerardo Iñiguez, and Márton Karsai. Multi-entrance phase transitions of complex contagion on multiplex networks. *to be published*, 2019.

[356] Samuel Unicomb, Gerardo Iñiguez, János Kertész, and Márton Karsai. Reentrant phase transitions in threshold driven contagion on multiplex networks. *Physical Review E*, 100(4):040301, 2019.

[357] Szabolcs Vajna, Bálint Tóth, and János Kertész. Modelling bursty time series. *New Journal of Physics*, 15(10):103023+, October 2013.

[358] Eugenio Valdano, Luca Ferreri, Chiara Poletto, and Vittoria Colizza. Analytical computation of the epidemic threshold on temporal networks. *Physical Review X*, 5(2):021005, 2015.

[359] Thomas W Valente. Social network thresholds in the diffusion of innovations. *Social networks*, 18(1):69–89, 1996.

[360] A. Vazquez. Impact of memory on human dynamics. *Physica A: Statistical Mechanics and its Applications*, 373:747–752, January 2007.

[361] Alexei Vázquez, Joao G. Oliveira, Zoltán Dezsö, Kwang-Il Goh, Imre Kondor, and Albert-László Barabási. Modeling bursts and heavy tails in human dynamics. *Physical Review E*, 73(3):036127+, March 2006.

[362] Alexei Vazquez, Balazs Racz, Andras Lukacs, and Albert-Laszlo Barabási. Impact of non-poissonian activity patterns on spreading processes. *Physical review letters*, 98(15):158702, 2007.

[363] Alessandro Vespignani. Predicting the behavior of techno-social systems. *Science*, 325(5939):425–428, 2009.

[364] Alessandro Vespignani. Modelling dynamical processes in complex socio-technical systems. *Nature physics*, 8(1):32, 2012.

[365] Peng Wang, Tao Zhou, Xiao-Pu Han, and Bing-Hong Wang. Modeling correlated human dynamics with temporal preference. *Physica A: Statistical Mechanics and its Applications*, 398:145–151, March 2014.

[366] Zhen Wang, Chris T Bauch, Samit Bhattacharyya, Alberto d'Onofrio, Piero Manfredi, Matjaž Perc, Nicola Perra, Marcel Salathé, and Dawei Zhao. Statistical physics of vaccination. *Physics Reports*, 664:1–113, 2016.

[367] Stanley Wasserman and Katherine Faust. *Social network analysis: Methods and applications*, volume 8. Cambridge university press, 1994.

[368] Duncan J Watts. A simple model of global cascades on random networks. *Proceedings of the National Academy of Sciences*, 99(9):5766–5771, 2002.

[369] Duncan J Watts and Peter Sheridan Dodds. Influentials, networks, and public opinion formation. *Journal of consumer research*, 34(4):441–458, 2007.

[370] Duncan J Watts and Steven H Strogatz. Collective dynamics of 'small-world'networks. *nature*, 393(6684):440–442, 1998.

[371] Victoria J. Wearmouth, Matthew J. McHugh, Nicolas E. Humphries, Aurore Naegelen, Mohammed Z. Ahmed, Emily J. Southall, Andrew M. Reynolds, and David W. Sims. Scaling laws of ambush predator waiting behaviour are tuned to a common ecology. *Proceedings of the Royal Society of London B: Biological Sciences*, 281(1782), 2014.

[372] Barbara Wejnert. Integrating models of diffusion of innovations: A conceptual framework. *Annu. Rev. Sociol.*, 28:297–326, 2002.

[373] Katrin Weller, Axel Bruns, Jean Burgess, Merja Mahrt, and Cornelius Puschmann. *Twitter and society*, volume 89. P. Lang, 2014.

[374] Lillian Weng, Alessandro Flammini, Alessandro Vespignani, and Fillipo Menczer. Competition among memes in a world with limited attention. *Scientific reports*, 2:335, 2012.

[375] M. S. Wheatland, P. A. Sturrock, and J. M. McTiernan. The Waiting-Time Distribution of Solar Flare Hard X-Ray Bursts. *The Astrophysical Journal*, 509(1):448, 1998.

[376] Stephen D White. Social media growth from 2006 to 2010. *All Things Marketing*, 2010.

[377] Ward Whitt. Approximating a point process by a renewal process, I: two basic methods. *Oper. Res.*, 30(1):125–147, 1982.

[378] D. S. Whote. Social media growth 2006 to 2012, 2013.

[379] Ian H Witten, Eibe Frank, Mark A Hall, and Christopher J Pal. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.

[380] Ye Wu, Changsong Zhou, Jinghua Xiao, Jürgen Kurths, and Hans Joachim Schellnhuber. Evidence for a bimodal distribution in human communication. *Proceedings of the national academy of sciences*, 107(44):18803–18808, 2010.

[381] Zhihao Wu, Giulia Menichetti, Christoph Rahmede, and Ginestra Bianconi. Emergent complex network geometry. *Scientific reports*, 5:10073, 2015.

[382] Jian Xu, Thanuka L Wickramarathne, and Nitesh V Chawla. Representing higher-order dependencies in networks. *Science advances*, 2(5):e1600028, 2016.

[383] Osman Yağan and Virgil Gligor. Analysis of complex contagions in random multiplex networks. *Physical Review E*, 86(3):036103, 2012.

[384] Shuicheng Yan, Dong Xu, Benyu Zhang, Hong-Jiang Zhang, Qiang Yang, and Stephen Lin. Graph embedding and extensions: A general framework for dimensionality reduction. *IEEE transactions on pattern analysis and machine intelligence*, 29(1):40–51, 2007.

[385] Ercan Yildiz, Asuman Ozdaglar, Daron Acemoglu, Amin Saberi, and Anna Scaglione. Binary opinion dynamics with stubborn agents. *ACM Trans. Econ. Comput.*, 1(4):19, 2013.

[386] Qian Zhang, Márton Karsai, and Alessandro Vespignani. Link transmission centrality in large-scale social networks. *EPJ Data Science*, 7(1):33, Sep 2018.

[387] Dawei Zhao, Lianhai Wang, Shudong Li, Zhen Wang, Lin Wang, and Bo Gao. Immunization of epidemics in multiplex networks. *PloS one*, 9(11):e112018, 2014.

[388] Kun Zhao, Márton Karsai, and Ginestra Bianconi. Entropy of dynamical social networks. *PloS one*, 6(12):e28116, 2011.

[389] Kun Zhao, Juliette Stehlé, Ginestra Bianconi, and Alain Barrat. Social network dynamics of face-to-face interactions. *Physical review E*, 83(5):056109, 2011.

[390] Xiaoxue Zhao, Takahiro Omi, Nanae Matsuno, and Shigeru Shinomoto. A non-universal aspect in the temporal occurrence of earthquakes. *New Journal of Physics*, 12(6):063010, 2010.

[391] W-X Zhou, Didier Sornette, Russell A Hill, and Robin IM Dunbar. Discrete hierarchical organization of social group sizes. *Proceedings of the Royal Society B: Biological Sciences*, 272(1561):439–444, 2005.

[392] Yong Zhuang, Alex Arenas, and Osman Yağan. Clustering determines the dynamics of complex contagions in multiplex networks. *Physical Review E*, 95(1):012312, 2017.